Psycho

Statistique en psychologie

Nathalie Rude
Olivier Retel

In press

STATISTIQUE EN PSYCHOLOGIE

Nathalie Rude et Olivier Retel

In press

IN PRESS EDITIONS
Serge et France Perrot, Éditeurs
12, rue du Texel - 75014 Paris
Tél. : 01 43 35 40 32
Fax : 01 43 21 05 00

Collection PSYCHO
Sous la direction de Aude Caria

Maquette
Marianne Adato

STATISTIQUE EN PSYCHOLOGIE
ISBN 2-912404-08-8
© 2000 IN PRESS ÉDITIONS
Toute représentation ou reproduction, intégrale ou partielle, faite sans le consentement des auteurs, ou de leurs ayants droit ou ayants cause, est illicite (loi du 11 mars 1957, alinéa 1er de l'article 40). Cette représentation ou reproduction, par quelque procédé que ce soit, constituerait une contrefaçon sanctionnée par les articles 425 et suivants du Code pénal.

LES AUTEURS

Nathalie Retel-Rude
est maître de conférence en Biomathématiques au Laboratoire de Biostatistiques pharmaceutiques de la Faculté de Médecine et de Pharmacie à l'Université de Franche Comté (Besançon) et membre du Laboratoire de psychologie sociale de l'Université Paris 5 (Boulogne).
Ses travaux portent sur l'approche psycho-sociale de la santé, sur la qualité de vie liée à la santé et sur la discordance soignant/soigné de la perception de la qualité de vie et des besoins associés.

Olivier Retel
est ingénieur à l'INSERM, Unité 444, Epidémiologie et sciences de l'information. Ses travaux portent sur la surveillance des maladies transmissibles en médecine générale et l'épidémiologie du VIH/VHC. Il exerce une activité de conseil en statistique et de développement d'outils de surveillance dans le domaine de la santé.

REMERCIEMENTS

Nous remercions cordialement Monsieur Philippe Bonnet, Maître de conférence en Statistique à l'UFR de Psychologie de l'Université Paris 5 ; Monsieur Bernard Cazelles, Maître de conférence en Biomathématiques à l'UFR de Biologie de l'Humain de l'Université Paris 7 ; Madame Sylvie Escolano-Rouyer, Ingénieur de Recherche à l'unité 436 de l'INSERM ; de même que Madame Aude Caria, psychologue clinicienne, responsable qualité au centre hospitalier Esquirol (Saint-Maurice), pour leur lecture attentive de cet ouvrage.

AVERTISSEMENT

Les statistiques et la psychologie sont des sciences en perpétuelle évolution. En même temps que les nouvelles recherches élargissent notre connaissance, il est nécessaire de faire progresser les méthodes. L'auteur et l'éditeur de ce livre ont vérifié la crédibilité de leurs sources afin de fournir des informations complètes et généralement en accord avec les normes en cours au moment de la publication. En revanche, en raison de la possibilité d'erreurs humaines ou d'évolution dans les sciences, ni l'auteur, ni l'éditeur, ni d'autres parties impliquées dans la préparation de ce livre ne garantissent que l'information contenue dans ce livre soit à tous égards exacte ou complète, et ils ne sont aucunement responsables de toute erreur, omission ou de résultats obtenus à partir de l'usage de ce document. Les lecteurs sont encouragés, et c'est aussi l'esprit dans lequel est rédigé l'ouvrage, à confirmer les informations de ce livre avec d'autres sources.

SOMMAIRE

CHAPITRE I - **S'intéresser aux statistiques**

1. Ayons une démarche analytique ! *15*

2. La statistique et les statistiques *18*
2.1. La statistique *18*
2.2. Les statistiques *19*

3. Intérêts de l'usage des statistiques en psychologie *20*
3.1. La psychologie, une discipline expérimentale *20*
3.2. Les différentes applications statistiques en psychologie *21*

4. Comment lire ce livre *22*

5. Détermination des tests statistiques *23*
5.1. Distinction entre variable qualitative ou quantitative *24*
5.2. Choix du test selon le type de variable *25*

CHAPITRE II - **Méthodologie - Applications**

1. Bien poser un problème de recherche *31*
1.1. Poser les hypothèses de recherche *31*
1.2. Comment s'orienter dans la littérature spécialisée ? *32*
1.3. Les bases de données en Psychologie *33*
1.4. Protocole d'étude *35*
1.5. Etablir un cahier des charges *43*

2. Acquisition des données *44*
2.1. Cahier d'observation ou questionnaire *44*
2.2. Type de questions *46*
2.3. Type de modalités de réponses *46*
2.4. Echelles de mesure *47*
2.5. Recueil des données *48*

3. Traitement des données *48*
3.1. Saisie des données *48*
3.2. Codage et recodage des données *50*

4. Analyse statistique des données *51*

5. Interprétation des résultats *52*

6. Procédure résumée des tests d'hypothèses *52*

7. Les logiciels statistiques *54*
7.1. Logiciels professionnels *55*
7.2. Logiciels sur ordinateur personnel *56*
7.3. Logiciels gratuits *57*
7.4. Sites Internet *58*

CHAPITRE III - **Des probabilités vers les statistiques**

1. Introduction aux probabilités *63*
1.1. Calcul combinatoire *63*
1.2. Calcul de probabilités *67*
1.3. Probabilités conditionnelles *71*
1.4. Indépendance *73*
1.5. Arbre des probabilités *74*
1.6. Théorème de Bayes *76*

2. Introduction aux statistiques *79*
2.1. Variable aléatoire, fonction de répartition *79*
2.2. Densité de probabilité d'une variable aléatoire continue *81*
2.3. Statistiques de position : Moyenne, Médiane, Mode *81*
2.4. Statistiques de dispersion : Etendue, Ecart-moyen, ... *86*
Ecart-type, Variance, Covariance

3. L'échantillonnage d'une population *91*

3.1. Randomisation 92
3.2. Echantillons appariés 93
3.3. Cas-témoin 93

4. Lois de probabilité 93
4.1. Loi binomiale 93
4.2. Loi de Poisson 94
4.3. Loi normale 95

5. Notion d'intervalle de confiance 103
5.1. Variance d'une moyenne 104
5.2. Intervalle de pari d'une moyenne 106
5.3. Intervalle de confiance d'une moyenne 107
5.4. Intervalle de pari d'un pourcentage 108
5.5. Intervalle de confiance d'un pourcentage 109

6. Exercices 110
6.1. Exercices corrigés 110
6.2. Exercices non corrigés 111
6.3. Solutions 113

CHAPITRE IV - **Statistiques descriptives : l'analyse préliminaire**

1. Description des données 119
1.1. Description de la population étudiée 119
1.2. Description des variables 120
1.3. Regroupement des données par classes 122
1.4. Tri à plat des données 123

2. Représentations graphiques des données 125
2.1. Diagrammes en bâtons 125
2.2. Histogrammes 127
2.3. Camemberts 129

CHAPITRE V - **Statistiques inférentielles**

1. Statistique de décision : Tests d'hypothèses 135
1.1. Test bilatéral 136
1.2. Test unilatéral à droite 138
1.3. Test unilatéral à gauche 139
1.4. Risques d'erreur et puissance d'un test 141

2. Tests paramétriques ou tests non paramétriques ? *143*

3. Test de Kolmogorov-Smirnov *144*
3.1. Cas de petits échantillons $5 < n_A < 30$ et $5 < n_B < 30$ *144*
3.2. Cas de grands échantillons $n_A \geq 30$ et $n_B \geq 30$ *146*

CHAPITRE VI - **Tests des différences entre variables qualitatives à deux ou n classes**

1. Données non appariées *151*
1.1. Test du χ^2 ..... *151*
1.2. Le χ^2 corrigé de Yates (tableau 2x2) *159*

2. Données appariées - χ^2 de Mac Nemar *162*

3. Récapitulatif *164*

4. Exercices *166*
4.1. Exercices corrigés *166*
4.2. Exercices non corrigés *167*
4.3. Solutions *169*

CHAPITRE VII - **Tests des différences entre variables quantitatives dans le cas de deux groupes**

1. Variables quantitatives dans le cas de deux groupes avec données non appariées *175*
1.1. Comparaison d'une moyenne observée à une moyenne théorique ($n \geq 30$) - test ε ou test z *175*
1.2. Comparaison d'une moyenne observée à une moyenne théorique ($n < 30$) - test t *179*
1.3. Comparaison de deux moyennes observées (avec n_A et $n_B \geq 30$) - test ε ou test z *182*
1.4. Comparaison de deux moyennes observées (avec n_A ou $n_B < 30$ et distributions normales de même variance) - test t *185*
1.5. Comparaison de deux lois observées (avec $n < 30$ et distribution non normale) - test de Mann et Whitney *188*
1.6. Comparaison de deux variances *201*

2. Variables quantitatives dans le cas de deux groupes avec données appariées *204*

2.1. Comparaison de deux moyennes observées (avec $n \geq 30$ et distribution normale) - test ε ou test z 204
2.2. Comparaison de deux moyennes observées (avec $n < 30$ et distribution normale) - test t 208
2.3. Comparaison de deux moyennes observées - test de Wilcoxon 211

3. Récapitulatif 215

CHAPITRE VIII - **Tests des différences entre variables quantitatives dans le cas de plus de deux groupes**

1. Analyse de variance à un facteur 225
Comparaison de plusieurs moyennes observées
1.1. Rôle du test 225
1.2. Principe du test 226
1.3. Procédure du test 229
1.4. Conditions d'application 231

2. Analyse de variance à deux facteurs croisés 234
2.1. Rôle du test 234
2.2. Principe du test à partir d'un exemple 235
2.3. Procédure du test 237
2.4. Conditions d'application 240

3. Comparaison de plusieurs moyennes - Test de Kruskall-Wallis 240
3.1. Rôle du test 240
3.2. Procédure du test 240
3.3. Cas des échantillons sans nœud 241
3.4. Cas des échantillons avec nœud 241
3.5. Exemples 242

4. Analyse de variance à deux facteurs avec répétition 247
4.1. Rôle du test 247
4.2. Procédure du test 247
4.3. Conditions d'application 250

5. Récapitulatif 251

CHAPITRE IX - Exercices des chapitres VII et VIII

1. **Exercices corrigés** *259*

2. **Exercices non corrigés** *264*

3. **Solutions** *268*

CHAPITRE X - Tests de liaison entre deux variables quantitatives

1. **Test d'indépendance entre deux variables quantitatives** *287*
   1.1. Adéquation entre le test et la question posée *287*
   1.2. Définition des termes corrélation et régression *288*
2. **La régression** *289*
   2.1. Indépendance entre deux variables quantitatives *289*
   2.2. Ajuster une droite au sens des moindres carrés *293*
   2.3. Analyse de l'équation de régression *293*
   2.4. Intervalles de confiance *298*
   2.5. Analyse des résidus *300*
3. **La corrélation** *303*
   3.1. Le coefficient de corrélation de Bravais-Pearson *304*
   3.2. Test et intervalle de confiance *307*
   3.3. La corrélation sur les rangs *308*
4. **Utilisation pertinente de la régression et/ou corrélation** *309*
5. **Exercices** *311*
   5.1. Exercices corrigés *311*
   5.2. Exercices non corrigés *312*
   5.3. Solutions *313*

GLOSSAIRE *317*

INDEX *321*

OUVRAGES DE RÉFÉRENCE *327*

TABLES STATISTIQUES *329*

CHAPITRE I

S'intéresser aux statistiques

1. **Ayons une démarche analytique !** *15*

2. **La statistique et les statistiques** *18*
   2.1. La statistique *18*
   2.2. Les statistiques *19*

3. **Intérêts de l'usage des statistiques en psychologie** *20*
   3.1. La psychologie, une discipline expérimentale *20*
   3.2. Les différentes applications statistiques en psychologie *21*

4. **Comment lire ce livre** *22*

5. **Détermination des tests statistiques** *23*
   5.1. Distinction entre variable qualitative ou quantitative *24*
   5.2. Choix du test selon le type de variable *25*

CHAPITRE I

S'intéresser aux statistiques

1. Ayons une démarche analytique !

Nous avons pu observer que la perspective de l'analyse statistique des données était une source d'appréhension pour de nombreuses personnes. Une barrière de chiffres reste parfois à franchir, et notre motivation au travers de ce livre est de faire apparaître la **réelle simplicité** de ce passage à l'acte statistique. Nous allons alors, ensemble, éclaircir les problèmes posés et comprendre les simples mécanismes à acquérir pour réaliser au mieux et agréablement les analyses statistiques des données observées.

■ **L'objectif de ce livre** n'est pas d'étudier les fondements mathématiques de l'analyse statistique, mais de **comprendre les principes qui permettent de réaliser correctement ces analyses**, quelle que soit la formation initiale, scientifique ou littéraire, du lecteur. Il est en effet parfois difficile pour certains de se pencher sur les démonstrations mathématiques qui conduisent aux formules statistiques courantes. Cela serait évidemment préférable pour acquérir une bonne compréhension des fondements de la statistique et de ses champs d'application. Mais c'est un travail considérable, voire trop difficile pour des non mathématiciens, et qui constitue une spécialité à part entière. Pour ceux qui souhaitent compléter leurs connaissances et/ou satisfaire leur curiosité pour les statistiques "sans s'arracher les cheveux" sur des formules mathématiques, nous avons tenté de faciliter l'approche des statistiques, de susciter leur curiosité à ce sujet afin de rendre cet enseignement agréable et profitable. L'exercice auquel nous nous sommes livrés est d'énoncer clairement le principe des tests de façon

parfois "trop simpliste" pour un statisticien, mais qui permettra au lecteur d'en comprendre l'esprit, et donc d'assimiler son utilisation. Pour ceux qui souhaiteront partir à la découverte des statistiques de façon plus approfondie, nous proposons des références d'ouvrages plus spécialisés dans la bibliographie, à la fin de ce livre.

Oublions donc tous les préjugés de complexité que nous pouvons avoir sur les statistiques, car ils ne sont pas fondés. Nous avons pu observer que certains étudiants sont perdus face aux statistiques, pensant ne pas avoir les bases nécessaires.

■ **Or il est simplement utile pour lire ce livre de :**

— savoir faire une addition, une multiplication ;

— savoir utiliser une machine à calculer ;

— savoir que | x |, se lit valeur absolue de x, et correspond à la valeur positive de x. Par exemple : | 1 | = 1 et | -1 | = 1

— savoir que $\sum_{i=1}^{n} x_i$ correspond à la somme de toutes les valeurs de la variable x de x_1 jusqu'à x_n, c'est-à-dire $\sum_{i=1}^{n} x_i = x_1 + x_2 + x_3 + \ldots + x_n$;

— savoir que nous notons les paramètres théoriques par des lettres de l'alphabet grec, et les statistiques observées ou estimées par des lettres de l'alphabet latin.

Aussi, pour une meilleure compréhension, nous convions le lecteur à **consulter entièrement le glossaire**, dès à présent, afin de se familiariser avec les termes employés par la suite.

Il est donc possible de comprendre la statistique sans grandes connaissances mathématiques. Par ailleurs, il paraît préférable d'approcher la statistique avec intérêt, dans le but d'une utilisation future. Gardons à l'esprit que la statistique est simplement un outil capable de valider ou d'invalider nos hypothèses.

■ **En ce sens, l'essentiel est de déterminer les tests statistiques à réaliser selon la question posée, selon les données en respectant les conditions d'application.** Or le choix du meilleur plan d'analyse n'est possible que si le principe des tests est compris. Afin d'optimiser cette approche, il est souhaitable, lors des examens, de pouvoir disposer de tous les documents nécessaires et possibles. Car l'essentiel de l'usage des statistiques, en psychologie et dans d'autres domaines, n'est pas de retenir par cœur les formules (pour les oublier l'année suivante), mais de savoir déterminer quel est le traitement statistique adapté à ses données. L'utilisation des logiciels statistiques doit être abordée de la même manière.

En statistique, la difficulté réside dans la définition des objectifs à atteindre, c'est-à-dire dans l'établissement du protocole d'étude, du plan d'analyse et du cahier d'observation. C'est pourquoi nous présentons les tests statistiques sous forme de procédure à suivre, compréhensible et à la portée de tous.

D'une manière pratique, il ne s'agit pas, une fois les données recueillies, de se demander quelle est la meilleure analyse statistique à mettre en œuvre.

■ **Il faut établir ce que nous voulons mettre en évidence en premier lieu.** En effet, avant le recueil des données, il est essentiel de déterminer toutes les variables à prendre en compte dans l'analyse statistique (voir paragraphe 5.1. de ce chapitre). Pour cela, il faut **au préalable poser les hypothèses à tester** dans le but de différencier des populations caractérisées par une ou plusieurs variables et/ou d'évaluer les liens entre ces variables selon les groupes observés (par exemple selon le sexe, l'âge, la situation géographique, la pathologie, les conditions d'expérience, etc.). De plus, il faut penser à contrôler les variables (ce que nous expliquons plus tard dans ce livre).

■ **Ainsi, le plan d'analyse statistique et le nombre de sujets nécessaire doivent être clairement définis avant le début de l'étude d'observation.** L'élaboration du questionnaire ou du cahier d'observation doit donc être basée sur l'analyse statistique qui en résultera. Bien entendu, le choix du type de variables qualitatives ou quantitatives (dont la distinction est présentée à la fin de ce chapitre) est également fonction de l'analyse statistique souhaitée et fonction des possibilités de recueil des données et bien entendu des objectifs. **C'est pourquoi la qualité des résultats d'une étude réside notamment dans la qualité du protocole d'étude** (ou protocole expérimental).

Une fois les données recueillies et présentées sous forme de tableaux ou de matrices, éventuellement saisies sur ordinateur, il ne reste plus qu'à suivre le plan d'analyse préétabli et à effectuer les tests statistiques qui ne constituent que des outils de synthèse de nos observations. Les conclusions des tests permettront alors d'interpréter les résultats. D'autres hypothèses et études complémentaires pourront par la suite être envisagées selon l'interprétation des résultats. La recherche est ainsi sans fin.

Enfin, insistons sur le fait que la formulation des conclusions et l'interprétation des résultats doivent toujours être replacées dans leur contexte, c'est-à-dire selon les caractéristiques de la ou des populations étudiées. Par exemple, des conclusions permettant d'affirmer qu'une expérience A est plus efficace qu'une expérience B doivent être détaillées

et comprendre une description complète de la population observée. En effet, des conclusions de cet ordre n'auront pas le même impact si l'étude est réalisée à partir de 20 sujets ou à partir de 1 000 sujets, selon la constitution des échantillons, selon la situation géographique ou l'origine culturelle des sujets, etc. De plus, des conclusions dans le sens d'un lien de causalité entre deux variables nécessitent des conditions d'expérience qu'il est essentiel de spécifier, par exemple une procédure de tirage au sort, et/ou une procédure de variation aléatoire des modalités d'une variable (contrôle des variables).

2. La statistique et les statistiques

L'origine des statistiques est allemande : "statistik" vient du terme *status* qui signifie état (état de l'art) en latin.

Plusieurs définitions sont données à "**la** statistique" :

— Etude des ensembles numériques et de leurs relations.

— Moyen d'obtenir des indications probables sur des ensembles imparfaitement connus.

— Science qui a pour objet le groupement méthodique, pour une époque déterminée, des faits sociaux qui se prêtent à une évaluation numérique. Méthode d'analyse et d'élaboration scientifique de ces faits.

Tandis que "**les** statistiques" constituent l'ensemble des renseignements numériques découlant des recensements de population, des données de registre d'état civil et d'enquête appropriées.

2.1. La statistique

La **"statistique"** représente un ensemble de données observées (numériques le plus souvent) dans une population ou une sous-population. Elle est un outil de synthèse des données observées. Elle est caractérisée par le recueil, le traitement et l'interprétation de ces données. La "statistique" est une méthode qui met en œuvre plusieurs "statistiques" caractérisant et résumant les données par dénombrement ou par répartition.

La statistique est employée dans différents domaines de recherche (psychologie, sociologie, biologie, épidémiologie, démographie, économie...) afin d'évaluer la variabilité des caractères étudiés d'un individu à l'autre (variabilité inter-individuelle), ou d'un individu d'un moment à l'autre (variabilité intra-individuelle).

2.2. Les statistiques

Les données observées sont recueillies par échantillonnage de larges ensembles de sujets, d'objets ou de mesures, le plus représentatif étant par tirage au sort. L'échantillon formé constitue alors un sous-ensemble de la population de référence (ou population parente), et les nombres résultant de la manipulation des données de l'échantillon sont appelés **statistiques**. Les statistiques sont les nombres qui caractérisent un échantillon et permettent de le décrire. Par exemple, la moyenne des observations est une des statistiques caractérisant l'échantillon. *En pratique, les statistiques calculées à partir d'un échantillon sont utilisées pour estimer les caractéristiques de la population de référence.* Les caractéristiques de la population sont appelées **paramètres**. Ainsi, la moyenne d'un échantillon est une statistique, et la moyenne d'une population est appelée paramètre (voir figure 1). Les statistiques (notées par des lettres de l'alphabet latin) ont pour objet d'estimer les paramètres (notés par des lettres de l'alphabet grec), c'est-à-dire que les statistiques permettent de déterminer les valeurs probables des paramètres inconnus.

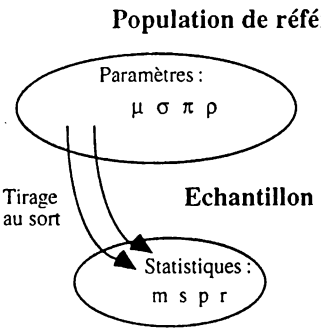


Figure 1

Notations	Théorique	Observé ou estimé
Moyenne	μ	m
Ecart-type	σ	s
Variance	σ^2	s^2
Proportion	π	p
Coefficient de corrélation	ρ	r

Les statistiques ont deux fonctions principales :

1) décrire la nature de l'échantillon, on parle alors de **statistiques descriptives** ;

2) inférer les caractéristiques d'une population (les paramètres) à partir de l'échantillon (des statistiques) ; c'est-à-dire déduire les paramètres de la population de référence à partir des statistiques estimées d'après les données observées. Il s'agit de déterminer l'impact des modalités de certaines variables sur d'éventuelles différences entre les individus observés. On parle alors de **statistiques inférentielles**. Une fois les paramètres inférés, on peut alors faire des études comparatives complémentaires et/ou étudier les liens entre les variables (ou caractères) étudiées.

3. Intérêts de l'usage des statistiques en psychologie

3.1. La psychologie, une discipline expérimentale

Les progrès de la psychologie reposent aussi bien sur des compétences cliniques, que sur une solide méthodologie de recherche et des analyses statistiques. Les outils cliniques tels que les tests psychologiques, la psychothérapie, la psychophysiologie et les neuro-sciences appliquées sont issus de recherches approfondies sur de nombreuses données recueillies à partir d'une multitude d'expériences. C'est pourquoi les techniques utilisées par les chercheurs en psychologie requièrent la connaissance des bases de la statistique ; que cela soit dans l'objectif de piloter des expérimentations, ou dans celui de comprendre et/ou d'affiner les résultats de recherches publiées dans la littérature.

Comme la plupart des recherches en sciences de la nature et de la vie ou en économie, de nombreuses recherches en psychologie introduisant des mesures ou des dénombrements nécessitent **le traitement statistique des données**. En revanche, contrairement à d'autres sciences, la mesure du comportement et notamment les variables comportementales issues des observations ne répondent pas facilement à des modèles théoriques. En effet, les variables comportementales sont si nombreuses et si différentes d'un sujet à l'autre que cela conduit, selon les phénomènes étudiés, à une très grande dispersion des mesures. C'est pourquoi, le plus souvent, les analyses statistiques permettent d'établir des **relations qualitatives**, plutôt que des relations quantitatives.

3.2. Les différentes applications statistiques en psychologie

Les méthodes statistiques sont utilisées dans les différentes disciplines de la psychologie afin de différencier les conduites déterministes des conduites aléatoires ou imprévisibles, afin d'étudier les liens entre les variables, et afin de vérifier les hypothèses des chercheurs. Nous citons ici quelques exemples afin d'illustrer l'intérêt de l'usage de la statistique en psychologie (se référer à l'ouvrage de Mariné et Escribe, 1998, pour plus de détails). De plus, les exercices proposés dans ce livre sont inspirés de recherches en psychologie.

- En **psychologie clinique**, en **psychopathologie**, comme en **psychologie de la santé**, l'usage de questionnaires diagnostiques standardisés, ou d'échelles de mesure telles que les échelles cliniques de l'anxiété, de la dépression, des symptômes psychologiques, ou encore de qualité de vie étant fréquent, l'usage des méthodes statistiques l'est aussi, notamment dans le but de valider ces échelles et d'interpréter leurs résultats, ou de comparer l'efficacité de traitements. Bien entendu, les démarches cliniques ne nécessitent pas de recourir à la statistique. En revanche, on peut envisager que l'étude de l'efficacité de différentes procédures cliniques peut recourir au traitement statistique, par exemple en comparant l'amélioration de l'état des patients selon différentes procédures, ou selon l'entreprise ou non d'une psychothérapie.
- En **psychologie cognitive et expérimentale**, de nombreuses mesures sont effectuées telles que celles de l'intelligence (Binet 1857-1911), de la mémoire (Hermann Ebbinghaus 1850-1909), des temps de réaction à des stimuli, ou des niveaux de perception à des stimulations sensorielles. Ces études et expériences nécessitent alors l'usage de la statistique afin notamment de comparer plusieurs groupes de sujets entre eux, selon une ou plusieurs variables. On s'intéressera également aux liens statistiques entre les variables, tels que le lien entre le quotient intellectuel d'enfants et celui de leurs parents.
- En **psychologie du travail**, on fait parfois appel à des tests d'évaluation des capacités ou des qualités relationnelles entrant en jeu dans les activités salariées ou à des tests psychotechniques. La validation de ces tests et l'analyse des observations en résultant nécessitent alors l'usage de méthodes statistiques.
- En **psychologie sociale**, on peut être amené à utiliser une méthodologie expérimentale, lors de l'observation des conduites de sujets soumis à des stimuli évoquant des situations sociales. Les interactions observées au sein d'un groupe de sujets pourront alors faire l'objet d'analyses statistiques dans le but de mettre en évidence des conduites caractéristiques. Par

exemple, afin de tester l'influence du jugement de compères de l'expérimentateur sur les réponses de sujets, lors de plusieurs épreuves de reconnaissance d'une forme présentée initialement, on peut être amené à comparer statistiquement la répartition des erreurs et des bonnes réponses d'un échantillon de sujets à la répartition théorique issue de recherches publiées. On peut se référer par exemple à l'expérience de Bruner et Goodman (1947) sur l'influence de la valeur sociale d'un objet sur sa perception.

■ En **psychologie du développement**, en **psychologie de l'enfant** et en **psychologie de l'éducation**, l'étude de l'évolution des comportements des sujets au cours du temps est courante. Par exemple, à partir d'observations chez plusieurs groupes d'enfants, on pourra être amené à comparer statistiquement la proportion des conduites motrices, cognitives, de jeu, de perception et/ou de langage selon l'âge.

■ En **psychologie différentielle**, on est souvent amené à faire des études comparatives des conduites de sujets observés dans une même situation, selon différentes cultures. Dans ce cas, on utilise alors des tests d'hypothèses afin de conclure ou non à des différences interindividuelles, interculturelles, voire intra-individuelles selon les situations, et donc intra-culturelles.

■ En **psychologie animale**, l'étude du comportement animal peut aussi conduire à des analyses statistiques, telles que l'étude de l'apprentissage chez des rats placés à plusieurs reprises dans un même labyrinthe comportant des impasses (Tolman, 1886-1951). Dans ce cas, on notera la proportion d'erreurs de parcours commises par chaque rat, chaque jour, ou le nombre moyen d'erreurs par jour pour différents groupes de rats. Ces proportions ou ces moyennes pourront alors être comparées afin de déterminer si elles diminuent ou non au fur et à mesure des expérimentations.

Bien entendu, les méthodes statistiques sont utilisées dans encore bien d'autres disciplines de la psychologie ; par exemple, en psychophysique (voir exemple de la combinaison de la loi Fechner et de la loi Bouguer-Weber présentée par Mariné et Escribe, 1998).

4. Comment lire ce livre

L'objectif de ce livre est de rendre compréhensibles les mécanismes d'analyse d'études d'observation et d'études expérimentales en psychologie, d'un point de vue scientifique.

■ **Toute étude d'observation est basée sur la réflexion préalable du plan d'analyse.** Il paraît nécessaire d'insister sur cette réflexion préalable, car nous avons rencontré de nombreux chercheurs faisant appel à nos

services en statistique, qui avaient recueilli, parfois sur plusieurs années, de multiples données qui s'avéraient inexploitables. Par exemple, lors de l'analyse des données, si la question du nombre de sujets nécessaires à prendre en compte dans l'étude n'a pas été posée au préalable, on peut se retrouver avec un nombre de variables étudiées plus important que le nombre de sujets, et par suite devant l'impossibilité d'analyser les liens entre ces variables.

MODE DE LECTURE DE CE LIVRE

La détermination des tests statistiques à effectuer lors d'une analyse étant un point clé et s'étant révélée un point obscur pour de nombreux étudiants, nous avons été particulièrement attentifs à sa présentation.

- C'est pourquoi, nous convions le lecteur à consulter dès à présent les arbres décisionnels figurant à la fin de ce chapitre (paragraphe 5.2.), afin d'avoir une vue d'ensemble des tests statistiques abordés dans ce livre (dans les chapitres VI à X), et afin d'avoir une vision claire et simplifiée de la démarche de détermination des tests statistiques à employer selon les variables étudiées (qualitatives et/ou quantitatives) et selon le type d'analyse (différences ou liaisons entre les variables).
- L'ensemble du chapitre II indique les bonnes pratiques méthodologiques à mettre en œuvre.
- Le chapitre III permet au lecteur de faire le lien entre les probabilités et les statistiques.
- Le chapitre IV présente les statistiques descriptives qui permettent de préparer les données à l'analyse statistique en coordonnant les informations recueillies. Les données sont alors visualisées et éventuellement regroupées. Puis, les chapitres V à X présentent les procédures de réduction des données à des statistiques permettant l'interprétation des résultats en évaluant les limites des conclusions.

5. Détermination des tests statistiques

La détermination des tests statistiques à mettre en œuvre dans l'analyse des données, c'est-à-dire à inscrire dans le plan d'analyse, est fonction du type d'analyse souhaitée et du type de données disponibles.

Une première analyse possible réside simplement dans la description des données (analyse descriptive, voir chapitre IV) par regroupement des individus en sous-populations.

L'analyse descriptive permet d'identifier les différences à mettre en évidence entre les groupes d'individus. Des études comparatives peuvent être envisagées, qu'il s'agisse de données observées qualitatives ou quantitatives. En revanche, les tests statistiques à utiliser pour analyser les données sont spécifiques du type de variable qualitative ou quantitative. C'est pourquoi il est essentiel de savoir bien les distinguer.

5.1. Distinction entre variable qualitative ou quantitative

Chaque individu ou unité statistique, appartenant à une population, peut être décrit par un ensemble de caractéristiques appelées variables ou caractères. **Ces variables peuvent être qualitatives ou quantitatives.**

L'objet de la statistique est de synthétiser des observations individuelles. On parle couramment de l'analyse des variables d'un échantillon, une variable étant un groupe de mesures ou une classification d'observations individuelles telles que l'âge, le sexe ou la taille. Les variables représentent des caractéristiques des sujets observés.

Nous allons souvent être amenés à distinguer les variables quantitatives des variables qualitatives. Par exemple, le sexe est une variable qualitative à deux états (ou modalités), l'âge est une variable quantitative.

■ **Les variables quantitatives sont caractérisées par des quantités numériques :** elles peuvent prendre des valeurs réelles sur un intervalle continu (variable quantitative continue) comme le temps, la taille, le poids ; ou constituer un intervalle d'entiers positifs (variable quantitative discrète) comme peut l'être l'âge compté en années sur l'intervalle [0 ; 130], le nombre d'enfants dans la famille, ou le quotient intellectuel.

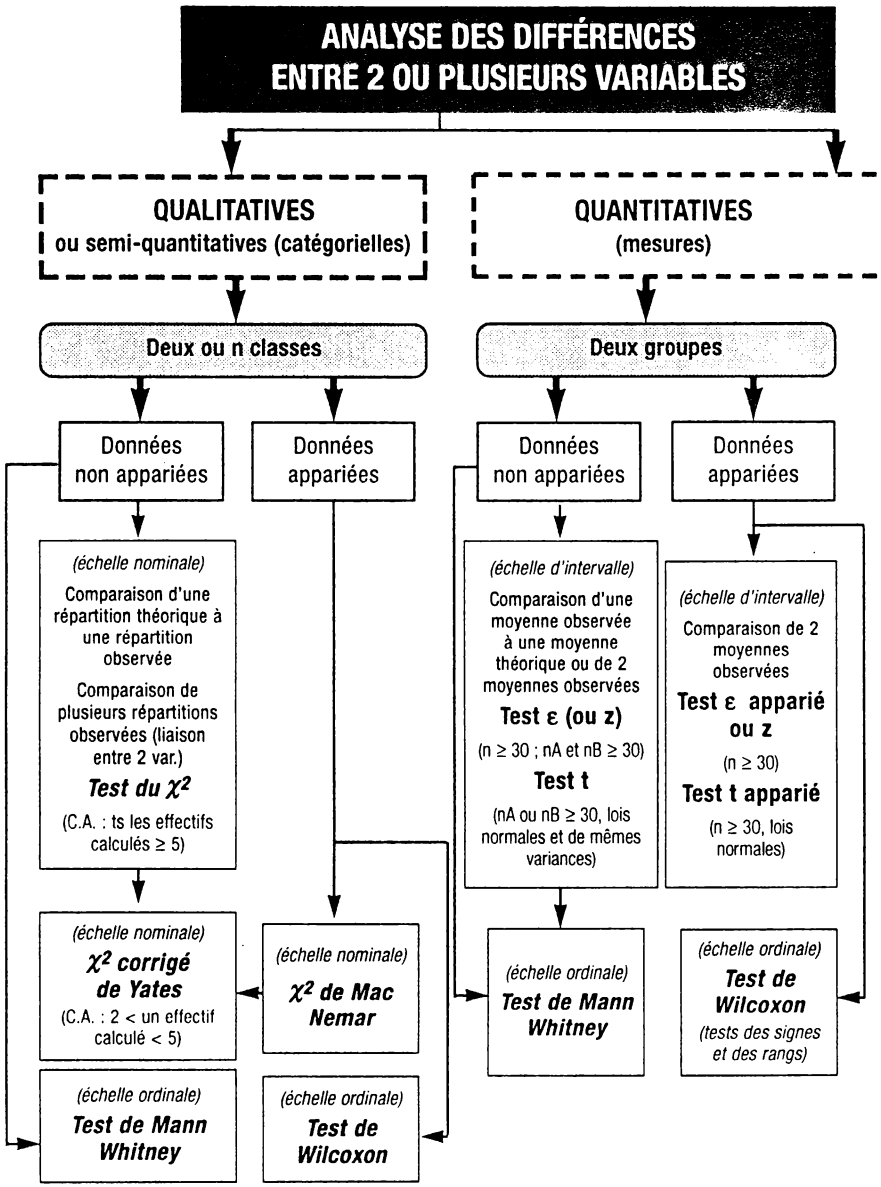
■ **Par opposition, les variables qualitatives sont définies par un ensemble de catégories,** où chaque observation appartient à l'une des catégories (ou modalités). Les variables qualitatives peuvent être classées en variables nominales (catégorie socioprofessionnelle) ou en variables ordinales (niveau de vie classé bas, moyen, élevé).

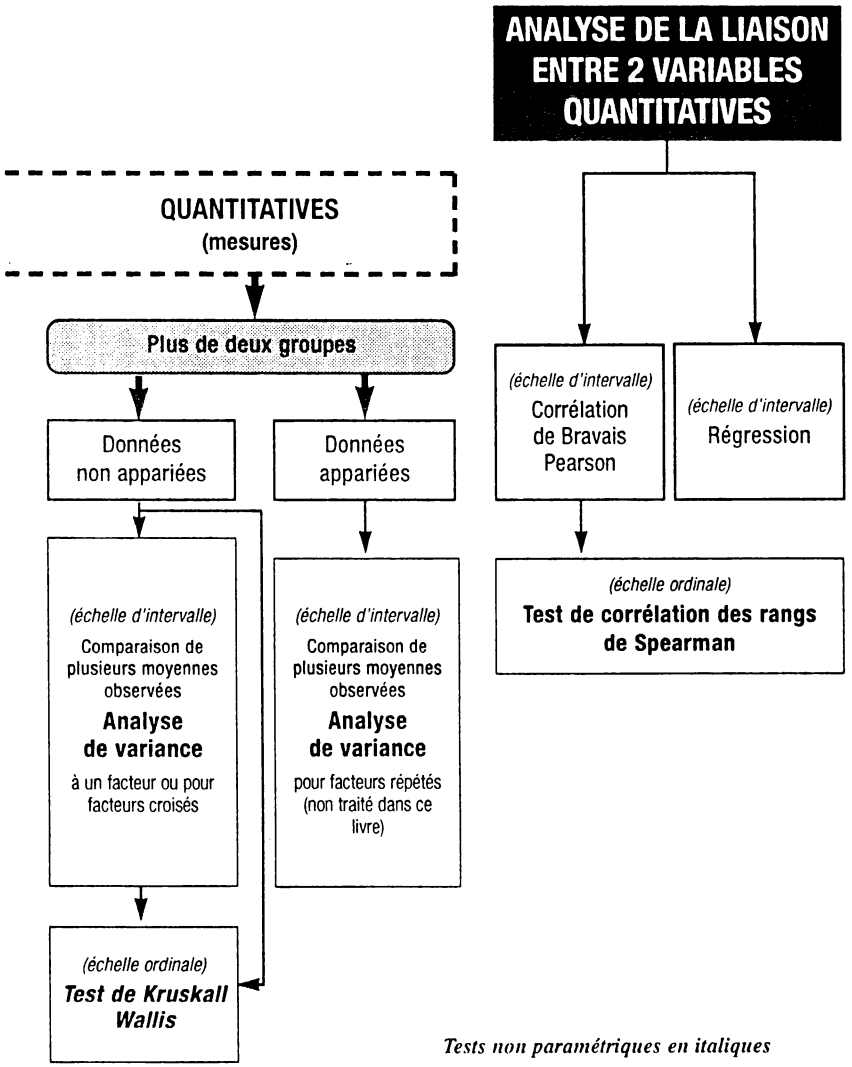
VARIABLES QUALITATIVES	VARIABLES QUANTITATIVES
sexe	performance ou efficacité mesurée
situation géographique	durée (en minutes, heures, jours, années)
catégorie socioprofessionnelle	note ou score
couleur des cheveux ou des yeux	mesure numérique
classe d'âge (semi-quantitative)	âge

■ **On peut transformer une variable quantitative en semi-quantitative (ou semi-qualitative)** en regroupant par exemple les individus par classes d'âge d'intervalles 10 ans. Si des individus sont recensés selon la classe d'âge à laquelle ils appartiennent, il s'agit alors d'une variable qualitative (puisqu'on ne peut distinguer les individus selon leur âge à l'intérieur d'une classe à moins de leur attribuer la valeur du centre de la classe d'appartenance). Il y a alors une "perte d'information" qui, le plus souvent, n'est pas souhaitable comme on le verra dans le chapitre X sur la régression et la corrélation. On effectue une telle transformation quand la variable n'a pas une distribution Gaussienne (normale) ou que le nombre d'individus est insuffisant ($n < 30$). Dans ce cas, le test de corrélation ne peut être appliqué, on réduit alors l'analyse à celle de variables qualitatives. ■

5.2. Choix du test selon le type de variable

Si les variables étudiées sont qualitatives (catégorielles), des analyses des différences entre des pourcentages ou proportions seront effectuées (voir chapitre VI). Si les variables étudiées sont quantitatives, alors des analyses des différences entre des moyennes pourront être effectuées (voir chapitres VII et VIII), de même que la mesure de la force des associations entre les variables (voir chapitre X). Ces analyses et tests associés au type de variable sont représentés dans l'arbre décisionnel suivant qui ne prétend pas être exhaustif mais introductif.


ANALYSE DES DIFFÉRENCES ENTRE 2 OU PLUSIEURS VARIABLES
QUALITATIVES
ou semi-quantitatives (catégorielles)
QUANTITATIVES
(mesures)
Deux ou n classes
Deux groupes
Données non appariées
Données appariées
Données non appariées
Données appariées
(échelle nominale)
Comparaison d'une répartition théorique à une répartition observée
Comparaison de plusieurs répartitions observées (liaison entre 2 var.)
Test du χ²
(C.A. : ts les effectifs calculés ≥ 5)
(échelle d'intervalle)
Comparaison d'une moyenne observée à une moyenne théorique ou de 2 moyennes observées
Test ε (ou z)
(n ≥ 30 ; nA et nB ≥ 30)
Test t
(nA ou nB ≥ 30, lois normales et de mêmes variances)
(échelle d'intervalle)
Comparaison de 2 moyennes observées
Test ε apparié ou z
(n ≥ 30)
Test t apparié
(n ≥ 30, lois normales)
(échelle nominale)
χ² corrigé de Yates
(C.A. : 2 < un effectif calculé < 5)
(échelle nominale)
χ² de Mac Nemar
(échelle ordinale)
Test de Mann Whitney
(échelle ordinale)
Test de Wilcoxon
(tests des signes et des rangs)
(échelle ordinale)
Test de Mann Whitney
(échelle ordinale)
Test de Wilcoxon



Tests non paramétriques en italiques

C.A. : Conditions d'application

CHAPITRE II

Méthodologie - Applications

1. Bien poser un problème de recherche *31*
1.1. Poser les hypothèses de recherche *31*
1.2. Comment s'orienter dans la littérature spécialisée ? *32*
1.3. Les bases de données en Psychologie *33*
1.4. Protocole d'étude *35*
1.4.1. Comment aborder un sujet de recherche sous l'angle scientifique et statistique ? *35*
1.4.2. Le plan type d'un protocole d'étude *36*
1.4.3. Comment mettre en œuvre une méthodologie permettant de produire des données exploitables ? *37*
1.4.4. Nombre de sujets nécessaires *38*
1.5. Etablir un cahier des charges *43*

2. Acquisition des données *44*
2.1. Cahier d'observation ou questionnaire *44*
2.2. Type de questions *46*
2.3. Type de modalités de réponses *46*
2.4. Echelles de mesure *47*
2.4.1. Echelles nominales ou qualitatives *47*
2.4.2. Echelles ordinales *47*
2.4.3. Echelles d'intervalles, numériques ou quantitatives *48*
2.5. Recueil des données *48*

3. Traitement des données *48*
3.1. Saisie des données *48*
3.2. Codage et recodage des données *50*

4. Analyse statistique des données *51*

5. Interprétation des résultats *52*

6. Procédure résumée des tests d'hypothèses *52*

7. Les logiciels statistiques *54*
7.1. Logiciels professionnels *55*
7.2. Logiciels sur ordinateur personnel *56*
7.3. Logiciels gratuits *57*
7.4. Sites Internet *58*

CHAPITRE II

Méthodologie - Applications

1. Bien poser un problème de recherche

1.1. Poser les hypothèses de recherche

Le choix de bonnes hypothèses de recherche repose notamment sur une bonne connaissance et une synthèse de la littérature dans le domaine étudié, puis sur l'originalité et la valeur heuristique (ajoutée) des hypothèses à vérifier. **La définition du problème posé** doit être très précise et détaillée afin de cibler au mieux la population à étudier et les variables qualitatives et quantitatives à prendre en compte dans le cahier d'observation (ou questionnaire). Une fois les variables identifiées, les **hypothèses à tester** pourront être précisées. Bien entendu, les variables identiquement établies dans la littérature pourront constituer une base de références pour la validation ou la simple comparaison des situations étudiées (voir le paragraphe 2. suivant, sur les outils de recherche bibliographique).

La mise en place d'**études pilotes préalables** ou études préliminaires (sur de petits échantillons de sujets) peut permettre de définir au mieux les hypothèses à tester, de préciser notamment les variables à étudier, et parfois de faire surgir de nouvelles variables à prendre en compte, ainsi que des difficultés de réalisation de l'enquête. Ces études préliminaires ne nécessitent pas forcément d'élaboration de protocole. En revanche, c'est en ayant déjà réfléchi à un protocole d'étude éventuel qu'une étude pilote peut être mise en place.

On entend par **protocole** le ou les documents présentant de façon détaillée la réalisation, la procédure, l'analyse, la bibliographie, et la publication des résultats d'une étude de données observées.

Par exemple, lors d'une étude sur la qualité de vie des personnes séropositives au Virus d'Immunodéficience Humaine (VIH) selon le "cocktail" de traitements administré (A ou B) les six derniers mois précédant l'étude, les hypothèses à tester seront : "le score moyen de qualité de vie ne diffère pas selon le type de traitement" versus "le score moyen de qualité de vie diffère selon le type de traitement" ou encore "le score moyen de qualité de vie sous le type de traitement A est supérieur au score moyen de qualité de vie sous le type de traitement B". Les hypothèses de base pourraient donc être :

$$\begin{cases} H_0 : \mu_A = \mu_B \\ \\ \text{contre} \quad \begin{cases} H_1 : \mu_A \neq \mu_B \\ \text{ou} \quad H_1' : \mu_A > \mu_B \\ \text{ou} \quad H_1'' : \mu_A < \mu_B \end{cases} \end{cases}$$

Notons que les tests unilatéraux sont fortement déconseillés par beaucoup de revues scientifiques à moins de réduire les seuils d'erreur.

1.2. Comment s'orienter dans la littérature spécialisée ?

Les sujets de recherche, même s'ils sont suscités par une motivation dans un domaine particulier, le sont aussi par la connaissance de la littérature spécialisée. En effet, la mise en œuvre d'études d'observation nécessite la connaissance de l'état de l'art du thème de recherche ; notamment afin de ne pas refaire ce qui a déjà été fait (donc voir ce qui est à faire) et afin d'obtenir des données de références qui permettront l'éventuelle validation des résultats obtenus. **Une recherche bibliographique** dans le domaine étudié permettra d'une part d'enrichir le plan d'analyse par la détermination éventuelle de nouvelles variables à prendre en considération, et d'autre part elle pourra apporter des éléments de référence dans l'objectif d'une validation des résultats et donc des modèles, échelles ou instruments de mesures élaborés. C'est pourquoi nous indiquons ici les principales bases de données bibliographiques disponibles.

Les différentes bases de données bibliographiques se caractérisent par la liste des revues qu'elles recensent, elle-même spécifique d'un domaine de recherche particulier. Ces bases permettent de retrouver, sur une période donnée (par exemple sur les cinq dernières années), l'ensemble des références de monographies et d'articles de périodiques du domaine étudié, défini par des mots-clés associés. La qualité d'une recherche bibliographique dépend avant tout du bon choix des **mots clés**. Afin de cibler au

mieux cette recherche, chaque base bibliographique dispose d'un répertoire des mots-clés faisant référence aux termes proches. Les mots clés étant souvent les mêmes d'une base à l'autre, une même recherche bibliographique peut facilement être faite sur plusieurs bases dans le but de recueillir le maximum de références du domaine étudié, notamment en psychologie de la santé ou en psychologie clinique.

1.3. Les bases de données en Psychologie

Les bases de données essentielles à consulter sur CD-ROM sont :

— **PsycLit** (ou Psychinfo), qui est "la" base de données en psychologie (toutes disciplines confondues) dans laquelle on retrouve la plupart des revues spécialisées en psychologie et également certaines du domaine médical ;

— **Pascal**, qui est une base de données en sciences, médecine et technologie, dans laquelle on retrouve aussi des articles en psychologie ;

— **Francis**, qui est une base de données multidisciplinaire en sciences humaines, sociales et en économie, issue du CNRS. Mais concernant la psychologie elle se limite à la psychologie du travail et aux sciences de l'éducation. On y retrouve cependant des revues dans d'autres disciplines telles que la *Revue française de psychanalyse*.

— **Medline**, qui recense notamment les revues spécialisées du domaine de la santé et comprend de nombreux articles en psychologie.

— puis **Cancerline**, qui recense bien entendu les articles spécialisés en cancérologie.

Aucune de ces bases de données n'est exhaustive des articles publiés en psychologie ; c'est pourquoi, il est souvent souhaitable d'en interroger plusieurs en même temps lors d'une recherche bibliographique sur un thème particulier, avec les mêmes mots clés ou d'autres selon la recherche.

Les bases de données bibliographiques peuvent être consultées dans certaines bibliothèques universitaires, malheureusement le plus souvent uniquement par ses membres ou ses étudiants.

Nous pouvons également citer la base de données bibliographiques multidisciplinaire espagnole, ISOC, disponible sur le CD-ROM CSIC (Consejo superior de investicationes cientificas) qui recense notamment des revues de psychologie et de sciences de l'éducation.

L'institut de l'information scientifique et technique (INIST) du CNRS est un service documentaire payant, donnant accès aux bases de données Francis et Pascal, soit par abonnement en contactant le service

client de l'INIST (2, allée du Parc Brabois - 54 514 Vandœuvre-les-Nancy), soit par minitel (au 08 36 29 36 01). Ce service permet également d'obtenir les copies des articles sélectionnés.

Le service payant d'**Information Médicale et Automatisée** (IMA) est un centre de documentation de l'INSERM (Institut National de la Santé et de la Recherche Médicale) qui permet d'obtenir les références d'articles sur un thème donné (ou par auteur) par courrier, par télécopie, par téléphone ou sur place sur rendez-vous (Information Médicale Automatisée, Centre de documentation de l'INSERM - Hôpital de Bicêtre - 80, rue du Général-Leclerc - 94 276 Kremlin-Bicêtre Cedex. Tél : 01 49 59 19 10 - Fax : 01 46 58 40 57). Ce service fournit une information complète et de qualité en sélectionnant les références à partir des bases de données les mieux adaptées au domaine qui vous intéresse (Medline, Pascal, Experta Medica, Biosis SCIsearch, Francis, Psychinfo, Cancerline, Chemical abstracts, Toxline, ...).

Le centre de documentation de la bibliothèque Henri Ey du centre hospitalier Sainte-Anne à Paris, riche dans le domaine de la psychiatrie, permet sur demande écrite (au Responsable de la bibliothèque médicale, Centre hospitalier Sainte-Anne, 1, rue Cabanis - 75 674 Paris Cedex 14) d'avoir accès à la bibliothèque et à la base de données Medline.

Par ailleurs, la plupart des bibliothèques universitaires ont accès au catalogue **MYRIADE** (sur CD-ROM) qui permet d'identifier la liste des lieux de consultation des revues, c'est-à-dire les bibliothèques ou les centres de documentation en France. Ce catalogue permet aussi d'avoir des informations pratiques sur les bibliothèques et les centres de documentation. Il suffit alors de contacter ces lieux afin de déterminer les bases de données bibliographiques dont ils disposent.

Un catalogue des livres disponibles en librairie, **ELECTRE** (sur CD-ROM), recense tous les livres publiés en France et les livres publiés en français à l'étranger. ELECTRE peut aussi être consulté sur minitel, 36 15 ELECTRE. De plus, le CD-ROM **Bibliographie Nationale Française** (BNF) décrit tous les ouvrages publiés en France depuis 1970 entrés à la Bibliothèque Nationale de France (BNF) par dépôt légal. L'équivalent britannique de ce CD-ROM, **British National Library** (BNL), est également disponible ; de même que son équivalent espagnol, **Bibliografía Nacional Española** (BNE).

1.4. Protocole d'étude

1.4.1. Comment aborder un sujet de recherche sous l'angle scientifique et statistique ?

La définition du sujet de recherche doit faire apparaître clairement le type de variables à étudier. Une fois l'état de l'art du domaine établi, la **définition de chaque variable à prendre en compte** doit être incluse dans le protocole d'étude. Puis la nature des variables entrant en jeu dans la confirmation ou l'infirmation des hypothèses de recherche, conduira à énoncer les **tests statistiques à réaliser**. Cela nécessite alors d'énoncer clairement les variables dépendantes et les variables indépendantes (voir glossaire).

Lors de l'élaboration du protocole d'étude, il s'agit d'anticiper les étapes et même les résultats afin de **déterminer le plan complet de l'étude**, du recueil des données jusqu'à leur analyse. Alors, et nous insistons tout particulièrement sur cette étape, le plan d'analyse prédéfini lors du protocole prédétermine le nombre de sujets à inclure dans l'étude (voir paragraphe 1.4.4. suivant, spécifique du calcul du "nombre de sujets nécessaires") et le plan expérimental s'il a lieu. Le plan d'analyse doit donc être réfléchi avant le recueil des données.

▩ **Afin d'établir un plan d'analyse statistique adapté à ses besoins, nous pouvons nous poser les questions suivantes :**

— **Quelles sont toutes les variables qualitatives et quantitatives prises en compte ?** Un tri à plat mettra en évidence et permettra de visualiser les caractéristiques des variables essentielles (variables expliquées, variables à expliquer, facteurs éventuels de confusion) et doit donc figurer dans le plan d'analyse. Une pré-enquête est le meilleur moyen de cerner les facteurs de confusion.

— **Quelles sont les distributions observées des variables considérées comme principales, puis des autres variables** (variables secondaires) ? Des représentations graphiques (courbes, histogrammes) permettront de visualiser les variables et peuvent donc figurer dans le plan d'analyse.

— **Quelles sont les hypothèses à tester ?** La liste des hypothèses à tester, distinguant les variables à considérer, conduira à énoncer les tests statistiques à réaliser, selon l'étude d'une liaison ou d'une différence entre les variables, et selon le type de données qualitatives ou quantitatives (voir les arbres de décision du chapitre I, paragraphe 5.2., sur la "détermination des tests statistiques").

— **Quelles sont toutes les hypothèses à tester ?** Des hypothèses complémentaires sont à envisager afin d'écarter au mieux les biais liés aux facteurs de confusion. Les tests correspondants figureront alors dans le plan d'analyse. Les conditions d'application des tests conduiront éventuel-

lement à d'autres tests et joueront un rôle important dans la détermination du nombre de sujets nécessaires à inclure dans l'étude.

Notons que le plan expérimental de l'étude détermine les possibilités d'interprétation des résultats. Pour cela les tirages au sort éventuels et le contrôle des variables sont primordiaux pour pouvoir conclure à des liens de causalité de façon scientifique.

1.4.2. Le plan type d'un protocole d'étude

Le plan type d'un protocole d'étude serait par exemple :

- Titre descriptif de l'étude
- Résumé (décrivant les objectifs, la procédure, les critères et le plan d'analyse de l'étude)
- Introduction
- Justification de l'étude
- But de l'étude et hypothèses de recherche
- Revue de la littérature
- Présentation des résultats d'études préliminaires éventuelles
- Sélection des sujets
- Nombre de sujets
- Critères d'inclusion
  - Caractéristiques des sujets
  - Critères diagnostiques
  - Consentement écrit (éventuellement)

  Critères de non inclusion
- Traitements, expérience ou mise en situation
  - Description des traitements ou de la situation
  - Description du matériel utilisé
  - Randomisation (tirage au sort)
  - Traitements ou expériences complémentaires (contre-indications)
- Description de l'étude
  - Méthodologie
  - Durée de l'étude
  - Déroulement de l'étude (calendrier de l'étude)
- Critères de jugement
  - Variables étudiées
    - Critère principal
    - Critères secondaires
  - Acceptabilité (éventuellement tolérance)
  - Observance (compliance au traitement ou l'expérience)
- Conditions expérimentales
  - Matériels et Techniques

Constitution des échantillons
- Recueil des données et suivi de l'étude
  Recueil des données
  Procédures de suivi de l'étude
- Procédures de saisie des données et vérification (contrôle de qualité)
- Effets indésirables ou inattendus
- Perdus de vue et sorties d'essai
  Sujets perdus de vue
  Sorties d'essai
- Considérations éthiques
- Méthodologie statistique — Plan d'analyse
- Budget de l'ensemble de l'étude
- Conditions de communication des résultats de l'étude
- Conditions d'archivage des données et des documents de l'étude
- Assurance de responsabilité civile (en cas de dangers éventuels entraînés par certaines procédures de l'étude envers les sujets observés, telles que des procédures thérapeutiques)
- Conformité à la loi informatique et libertés et anonymat des données (éventuellement)
- Amendements (éventuelles modifications du protocole initial)
- Références bibliographiques
- Annexes
  Liste des centres ou des lieux de recueil des données (dans le cas d'études multicentriques)
  Plan expérimental
  Cahier d'observation
  Informations aux sujets observés
  Formulaire du consentement éclairé écrit (éventuellement)
  Police d'assurance responsabilité civile (dans le cas d'essais thérapeutiques)

1.4.3. Comment mettre en œuvre une méthodologie permettant de produire des données exploitables ?

Il est nécessaire, avant le début de l'étude, de vérifier que tous les problèmes à venir ont été envisagés. On oublie trop souvent que le statisticien n'est pas seulement celui qui vient analyser les résultats, mais aussi celui qui prévoit les problèmes, aide à mettre en place l'étude et à la rendre exploitable.

Pour cela, on peut déjà vérifier les points suivants avant de lancer une étude :

– Réfléchir à l'acquisition optimale des données afin de réduire le nombre de données manquantes et de perdus de vue.

– Poser toutes les hypothèses nulles (H_0) sous forme statistique.
– Contrôler les variables.
– Déterminer les tests appropriés aux hypothèses posées et au type de données (quantitatives ou qualitatives).
– Calculer le nombre de sujets nécessaires à l'analyse des données (voir paragraphe suivant).

1.4.4. Nombre de sujets nécessaires

Il s'agit, avant d'entreprendre une étude, de calculer le nombre de sujets n qu'il faudra recruter dans l'échantillon au risque d'erreur α. Cela permet de savoir à l'avance si l'étude est réalisable et combien elle devrait coûter. Pour calculer ce nombre de sujets nécessaires, il faut bien savoir ce que l'on veut montrer et quel test statistique sera effectué dans l'analyse des résultats.

Voyons la procédure à partir des deux exemples suivants.

• Exemple 1 :

Supposons que sur l'ensemble des lycées d'une ville, il y ait systématiquement 60 % des élèves qui réussissent leur baccalauréat chaque année sans que l'on ait pu montrer une différence selon les lycées, le sexe ou l'âge. Une équipe de psychologues met alors en place une assistance pédagogique aux élèves de terminale de cette ville pour les aider à mieux appréhender les examens. Elle veut évaluer l'impact de cette assistance en milieu d'année en faisant passer un bac blanc à un groupe d'élèves. Combien faudra-t-il d'élèves dans cet échantillon pour montrer, au risque $\alpha = 5$ %, que le pourcentage de réussite p au bac blanc est significativement plus fort que le pourcentage théorique $\pi = 60$ % ?

Le test à effectuer est un test unilatéral à droite comme cela est décrit dans le chapitre V. En effet on ne s'attend pas à ce que le pourcentage de réussite soit inférieur à 60 %, alors même que les élèves auront bénéficié d'une aide pédagogique. On peut réécrire cela sous une forme plus statistique en faisant les hypothèses :

$$\begin{cases} H_0 : \pi = 60 \\ H_1 : \pi > 60 \end{cases}$$

avec π, le pourcentage théorique de réussite et p, le pourcentage observé à partir de l'échantillon envisagé.

L'intervalle de confiance (voir chapitre III) $[p_{inf} ; p_{sup}]$ à 95 % (avec un risque d'erreur $\alpha = 5$ %) d'un pourcentage est :

$$[p - \varepsilon_{5\%} \sqrt{\frac{pq}{n}} ; p + \varepsilon_{5\%} \sqrt{\frac{pq}{n}}].$$

L'équipe veut donc montrer que la valeur minimum p_{inf} de l'intervalle de confiance de l'échantillon est supérieure à $\pi = 60$ %, autrement dit :

$p_{inf} > \pi$, soit $p - \varepsilon_{5\%}\sqrt{\frac{pq}{n}} > \pi$.

On peut alors transformer cette relation pour en déduire n, avec $\varepsilon = 1{,}645$ au seuil $\alpha = 5$ % (puisqu'il s'agit d'un test unilatéral) :

$$p - 1{,}645\sqrt{\frac{pq}{n}} > \pi$$

$$\Leftrightarrow -1{,}645\sqrt{\frac{pq}{n}} > \pi - p$$

On multiplie par - 1, donc on inverse le sens de l'inégalité :

$$\Leftrightarrow 1{,}645\sqrt{\frac{pq}{n}} < p - \pi$$

On élève au carré chaque côté de l'inégalité :

$$\Leftrightarrow 1{,}645^2\,\frac{pq}{n} < (p-\pi)^2$$

$$\Leftrightarrow 1{,}645^2\,\frac{pq}{(p-\pi)^2} < n$$

On sait que p + q = 1, soit q = 1 - p, d'où la formule :

$$\Leftrightarrow n > \frac{2{,}706\,p(1-p)}{(p-\pi)^2} \qquad \text{équation (1)}$$

Cette inégalité permet de calculer n à partir des valeurs de π et p. On connaît $\pi = 0{,}6$, donc si l'on fixe un pourcentage attendu de réussite p à l'examen blanc, on peut alors déduire le nombre d'élèves minimum n à prendre en compte dans l'échantillon.

Supposons que l'on pense le programme d'assistance suffisamment efficace pour que 70 % des élèves réussissent (p = 0,7) alors on trouve (n'oubliez pas d'arrondir au nombre supérieur uniquement à la fin du calcul) :

$$n > \frac{2{,}706 \times 0{,}7 \times (1-0{,}7)}{(0{,}7-0{,}6)^2}$$

$$\Leftrightarrow n > \frac{2{,}706 \times 0{,}7 \times 0{,}3}{0{,}1^2}$$

$$\Leftrightarrow n > \frac{2{,}706 \times 0{,}21}{0{,}01}$$

$$\Leftrightarrow n > 56{,}83$$

Donc si on fait passer le bac blanc à n = 57 élèves et que p = 70 % d'entre eux réussissent, on aura montré qu'avec le programme pédagogique, la

réussite est plus importante que celle observée généralement dans cette ville de façon significative avec une risque d'erreur $\alpha = 5\ \%$.
On peut recalculer l'intervalle de confiance $[p_{inf} ; p_{sup}]$ à 95 %, pour bien vérifier que p_{inf} est supérieur ou égal à $\pi = 60\ \%$:

$$p_{inf} < p < p_{sup}$$

soit, $$p-1{,}645\sqrt{\frac{pq}{n}} < p < p+1{,}645\sqrt{\frac{pq}{n}}$$

▩ ***Application numérique*** *:*

$$0{,}7-1{,}645\sqrt{\frac{0{,}7\times 0{,}3}{57}} > p > 0{,}7+1{,}645\sqrt{\frac{0{,}7\times 0{,}3}{57}}$$

$$\textit{soit},\quad 0{,}7-1{,}645\frac{\sqrt{0{,}21}}{57} > p > 0{,}7+1{,}645\frac{\sqrt{0{,}21}}{57}$$

$$\Leftrightarrow \quad 0{,}7-0{,}1 > p > 0{,}7+0{,}1$$

$$\Leftrightarrow \quad 0{,}6 > p > 0{,}8$$

Et p est donc bien supérieur à 0,6 avec un risque d'erreur de 5 %.
Mais si au final seulement 65 % de ces élèves réussissent au bac blanc au lieu de 70 %, alors le test ne sera pas concluant, c'est-à-dire que l'on ne pourra pas conclure à une différence significative même si elle existe. On peut donc recalculer comme précédemment combien il faudrait d'élèves pour montrer une différence si seulement p = 65 % réussissent. Si ce nombre n n'est pas trop grand, on pourra alors se baser sur ce nouvel échantillon pour éviter le risque de ne pouvoir conclure à une différence.
Pour p = 65 %, on a :

$$n > \frac{2{,}706\times 0{,}65\times(1-0{,}65)}{(0{,}65-0{,}6)^2}$$

$$\Leftrightarrow \quad n > \frac{2{,}706\times 0{,}65\times 0{,}35}{0{,}05^2}$$

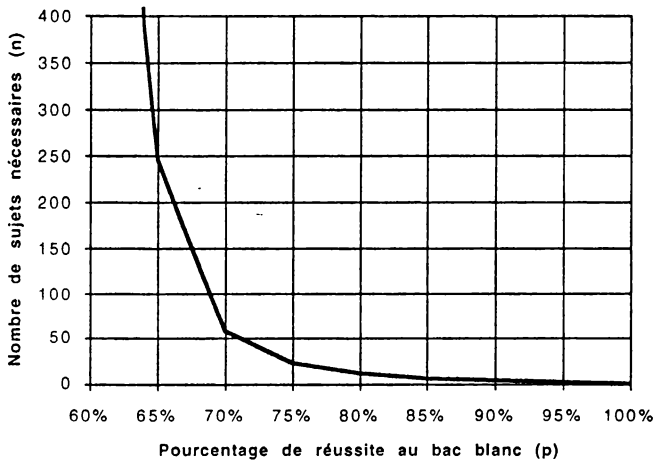
$$\Leftrightarrow \quad n > \frac{2{,}706\times 0{,}2275}{0{,}0025}$$

$$\Leftrightarrow \quad n > 246{,}25$$

On a besoin d'un échantillon d'au moins 247 élèves pour conclure à l'efficacité de l'assistance pédagogique si p = 65 %.

▩ ***Représentation graphique :***

Comme on connaît la relation entre n et p, on peut tracer le graphique suivant à partir de l'équation (1).



L'équation (1) nous permet de constater que la taille de l'échantillon doit être de 6 438 élèves pour mettre en évidence une différence de +1 % (p = 61 %), 1 594 si le pourcentage de réussite à mettre en évidence est p = 62 % (+2 %), 701 si p = 63 %, 390 si p = 64 % ; la taille de l'échantillon se réduit à 23 élèves pour mettre en évidence un pourcentage de réussite de +25 % (p = 75 %). Ce pourcentage se réduirait à 6 élèves pour p = 85 %, mais entre 85 % et 100 % de réussite, les conditions d'applications ne sont pas remplies puisque np < 5.

▩ ***Remarques :***

Le choix de la taille de l'échantillon peut se décider sur la base de cette figure. Cela dépend de l'objectif (est-ce que 65 % de réussite est une amélioration suffisante pour que le programme d'assistance pédagogique soit mis en place ? ou est-ce qu'il faut au moins une amélioration de 10 % ?) et du coût (est-ce que le budget de l'étude permet de faire passer un bac blanc à 247 élèves ?). Dans tous les cas, on rappelle qu'il y aura un biais de recrutement si ces élèves sont sélectionnés dans un même lycée : *il faut tirer au sort les sujets parmi l'ensemble des élèves de terminale de la ville.* Finalement, le psychologue-statisticien de l'équipe ne donne pas une seule réponse sur la taille que doit avoir l'échantillon, il permet à l'équipe de décider au mieux en précisant le rapport coût/efficacité.

Nous verrons que, pour la plupart des tests, la taille n de l'échantillon intervient dans la formule. Donc comme dans l'exemple ci-dessus, il est possible de faire une hypothèse sur le résultat attendu du test (ici nous avions fait une hypothèse sur le pourcentage p attendu) et de déduire la taille n de la formule d'après le test utilisé.

Il peut y avoir plusieurs tailles d'échantillons n à calculer : si l'on veut comparer les Q.I. des élèves de deux villes à partir de deux échantillons de taille n_1 et n_2, la variable étudiée étant quantitative, le test à effectuer sera une comparaison de deux moyennes (voir chapitre VII). Dans ce cas, on fera une hypothèse sur la moyenne et la variance du Q.I. attendu dans chaque ville et on choisira en général $n_1 = n_2$ (le test est plus puissant). La seule valeur inconnue est alors $n = n_1 = n_2$, qu'il faut déduire de la formule du test de comparaison de deux moyennes.

• *Exemple 2 :*

On suppose que les élèves de cette même ville qui réussissent leur baccalauréat ont un Q.I. plus élevé que ceux qui échouent. On sait par les redoublants que ceux qui échouent ont un Q.I. de moyenne $m_1 = 120$ avec une variance de 900. Combien doit-on prendre en compte d'élèves qui ont réussi leur baccalauréat, pour mettre en évidence une différence significative au risque $\alpha = 5\ \%$ si le Q.I. obtenu est de moyenne m_2 d'au moins 125 et de même variance ?

On sait que la probabilité pour que la moyenne d'un échantillon soit comprise dans cet intervalle est de 95 % :

$$m_1 < m_2 + \varepsilon \frac{s}{\sqrt{n}}$$

$$\Leftrightarrow \quad m_1 - m_2 < \varepsilon \frac{s}{\sqrt{n}}$$

$$\Leftrightarrow \quad \sqrt{n} > \varepsilon \frac{s}{m_1 - m_2}$$

$$\Leftrightarrow \quad \boxed{n > \varepsilon^2 \frac{s^2}{(m_1 - m_2)^2}}$$

Application numérique :

$$n > 1{,}645^2 \frac{900}{5^2} \quad \text{avec } \varepsilon = 1{,}646 \text{ pour } \alpha = 5\ \%$$

$$n > 2{,}706 \times 36$$

$$n > 97{,}42$$

Si l'échantillon est d'au moins 98 élèves et que ceux-ci aient un Q.I. moyen de 125, alors on peut mettre en évidence une différence significative de Q.I. entre ceux qui réussissent et ceux qui échouent au baccalauréat. Evidemment, certaines études ont montré que lorsque l'on communique les Q.I. des élèves à un professeur, son attitude est différente et leur réussite n'est pas la même : il s'agit alors d'un biais dans l'enquête. On parle d'effet pygmalion (Rosenthal et Jacobson, 1971), où des questions d'éthique se posent d'un point de vue méthodologique (dans le cas d'effet pygmalion négatif).
Ce calcul peut être plus long s'il existe une différence de Q.I. en fonction du lycée, car il faut alors préparer des échantillons dans chaque lycée pour pouvoir conclure à une différence entre villes. Le test à effectuer est une analyse de variance à deux facteurs (facteur ville, facteur lycée) et il faut calculer les tailles d'échantillons pour chaque lycée des deux villes.
Notons qu'il existe des logiciels pour micro-informatique, spécialisés dans les calculs de nombres de sujets nécessaires.

1.5. Etablir un cahier des charges

Dans le but d'organiser au mieux l'étude d'observation, il est souhaitable de rédiger un cahier des charges qui présentera point par point les étapes à réaliser pour mettre en place l'étude. Ce cahier des charges permettra notamment de définir les tâches de chacun des acteurs de l'étude, les contacts à établir ou le détail du suivi de l'étude, par exemple lors d'un monitoring ou d'études longitudinales.

Il s'agit d'organiser le plan pratique de l'étude, de la recherche bibliographique jusqu'à la présentation des résultats au public ou éventuellement dans une revue spécialisée.

Ce cahier des charges comprend notamment :

- identification des investigateurs,
- organisation des réunions avec les investigateurs,
- calendrier de l'évolution du protocole et de l'étude en elle-même,
- compte rendu détaillé de l'évolution de l'étude, des entretiens et des réunions,
- rapports des visites dans les différents centres de recueil des données,
- définition et description du matériel employé,
- vérification de la bonne utilisation du matériel ou des produits employés,
- vérification des compétences des investigateurs,
- vérification de la conformité du déroulement de l'étude au protocole,
- vérification du consentement éclairé écrit des sujets selon les besoins de l'étude,

- vérification de la correspondance et des archives éventuelles,
- compte rendu de la correspondance et des communications téléphoniques,
- instructions relatives au remplissage du cahier d'observation,
- observations et questions concernant le cahier d'observation,
- vérification du cahier d'observation afin d'apprécier la qualité du recueil des données,
- vérification du code d'identification et du respect de l'anonymat,
- mise en évidence des données manquantes ou incohérentes,
- description des perdus de vue,
- détermination du motif de fin d'étude,
- recueil des informations générales de l'étude (nombre de sujets, acceptabilité),
- description des mesures à prendre en cas d'effets inattendus, modification du protocole,
- description des effets indésirables ou inattendus,
- comptabilité.

2. Acquisition des données

2.1. Cahier d'observation ou questionnaire

Le cahier d'observation ou questionnaire, qui sera le support papier (ou informatique si la saisie est instantanée) de recueil des informations ou données souhaitées, sera le plus souvent partitionné pour davantage de clarté. Le cahier d'observation est composé de plusieurs parties, éventuellement de plusieurs questionnaires spécifiques.

Un cahier d'observation peut être constitué des différentes parties suivantes, par exemple pour ce qui concerne une étude de qualité de vie chez des patients dépressifs :

— identification du sujet,
— différentes échelles de mesure du déficit cognitif,
— une échelle de qualité de vie spécifique de la dépression,
— une échelle diagnostique psychiatrique,
— évaluation du stade de la pathologie et des symptômes associés,
— définition du ou des traitements du sujet les six derniers mois,
— revue des effets secondaires du ou des traitements,
— mesures de l'efficacité des différentes procédures thérapeutiques.

Le plus souvent, la première partie concerne les variables descriptives, distinctives des individus, c'est-à-dire les variables d'identification des sujets.

On peut s'intéresser par exemple, selon l'étude, aux variables suivantes :

- l'âge,
- la situation géographique,
- le statut socio-économique,
- le niveau d'étude,
- le statut marital,
- le nombre d'enfants,
- la pathologie actuelle,
- ...

Le nom est le plus souvent remplacé par un numéro d'observation, afin de respecter l'anonymat des sujets inclus dans l'étude.

■ **Une attention toute particulière doit être attachée à la forme du cahier d'observation, après avoir étudié le fond.** En ce qui concerne le fond, les variables figurant dans le cahier d'observation ont été déterminées au préalable lors de la constitution du protocole d'étude. Concernant la forme, le cahier d'observation pouvant être rempli par différentes personnes, que cela soit intégralement par des expérimentateurs et/ou partiellement par les sujets observés eux-mêmes, il doit être facile et pratique à remplir. Cela suppose de constituer une **maquette agréable à l'œil**, en évitant de concentrer trop d'informations au même endroit. L'expérimentateur comme le sujet observé, amenés à remplir ces questionnaires ou cahiers d'observations, ne doivent pas être fatigués, gênés, ou dérangés par une présentation confuse. L'utilisation d'encadrements, de couleurs, de flèches ou de caractères gras peut permettre une meilleure clarté de l'ensemble du document.

Lorsque les sujets sont sollicités pour certaines parties du cahier d'observation, un système de feuillets détachables peut être envisagé pour une meilleure gestion du recueil des données. Les cahiers d'observation pouvant être remplis par différentes personnes, les questions ou items doivent être rédigés simplement, c'est-à-dire de façon compréhensible par tous. En pratique, les questions doivent être comprises par n'importe quel adolescent de 14 ans. Le **vocabulaire utilisé** doit alors s'approcher au mieux de celui employé par les personnes susceptibles de répondre au questionnaire. De plus, les items ne doivent pas comprendre de double négation ; et doivent clairement définir une période de temps limitée (par exemple "Au cours des 2 dernières semaines, ..." ou "Actuellement, ...").

Particulièrement, dans le cas où les sujets observés doivent répondre eux-mêmes aux items, **le temps de réponse total ne doit pas être trop long** ; un maximum de vingt minutes paraît raisonnable. Bien entendu, certaines études demanderont davantage d'attention de la part des personnes observées et devront donc s'attacher tout particulièrement à présenter un mode de

réponse facile. Dans ce cas, le mode de réponse dichotomique (Oui/Non) sera privilégié. Le type de question doit alors être étudié lors de la constitution du cahier d'observation, en fonction des hypothèses et du protocole.

Notons bien que les modalités de réponse proposées dans les cahiers d'observation présageant l'analyse des données, elles doivent être définies lors de la constitution du plan d'analyse statistique.

2.2. Type de questions

On parle de *question ouverte,* quand celle-ci permet une réponse libre c'est-à-dire que le sujet observé répond à sa façon, à son rythme, avec son propre vocabulaire. L'usage des statistiques est limité lors de l'analyse de questions ouvertes. En revanche, lors de l'analyse de contenu des discours émis, on peut par exemple être amené à comptabiliser les occurrences des thèmes abordés, des temps de conjugaison employés, des négations, des affirmations, afin de déterminer les caractéristiques des réponses des sujets à partir des statistiques obtenues. Des logiciels d'analyse de contenu des discours permettent facilement de calculer ces statistiques (SPADT, Alceste...).

On parle de *question fermée* quand celle-ci est limitée par des modalités de réponses. Les différentes modalités de réponses définiront le type d'échelle de mesure.

2.3. Type de modalités de réponses

Les **réponses libres** issues de questions ouvertes doivent être rédigées clairement par le répondant si l'analyse prévoit une analyse de contenu avec saisie informatique. Pour faciliter cette clarté dans le cas d'un questionnaire sur papier, il paraît nécessaire d'une part de laisser suffisamment d'espace de réponse et, d'autre part de guider la rédaction par des lignes en pointillé. Dans le cas d'un entretien enregistré puis saisi sur ordinateur, bien entendu il n'est pas souhaitable de parler en même temps que le répondant (à moins de souhaiter le déstabiliser), il est préférable de laisser un court temps de silence entre chaque question, pour une meilleure qualité de saisie.

Les modalités de **réponses dichotomiques** proposent deux modalités (exemple : oui/non).

Les questions à **choix multiples** proposent deux ou plus de deux modalités. Il est souhaitable d'éviter un nombre impair de modalités afin d'éliminer des réponses "moyennes" sans prise de position.

Le choix de plus de deux modalités de réponses favorise une pondération des modalités de réponses plus nuancée, c'est-à-dire plus sensible. C'est le cas des **échelles de Likert**.

Les réponses recueillies à partir d'**échelle visuelle analogique** (EVA) linéaire, c'est-à-dire à partir d'une échelle de valeur constituée par une ligne de 10 centimètres (le plus souvent) indiquant aux deux extrémités des attitudes "extrêmes" ou opposées, facilite la pondération des items des questionnaires. Par exemple, pour la question "indiquez par une croix votre sensation de douleur sur l'échelle visuelle analogique suivante", on placera en dessous ou à côté de cette question la ligne suivante :

Aucune douleur_____ Douleur insupportable

Dans ce cas, le poids accordé à cette question pourra être une combinaison de la mesure marquée d'une croix par les sujets interrogés.

2.4. Echelles de mesure

2.4.1. Echelles nominales ou qualitatives

Dans le cas d'échelles nominales ou qualitatives, aucune structure mathématique ne peut relier les différentes modalités de l'échelle. Par exemple, le caractère qualitatif "couleur des yeux" peut prendre différentes modalités : verts, bleus, marrons ; pour lesquelles **aucune relation d'ordre ne peut être proposée** ; de même que pour les caractères "sexe" (masculin/féminin), "catégories socioprofessionnelles", "statut marital", "situation géographique", "cas contrôle ou témoin". La modalité "autre réponse" ou "ne sait pas" peut être adjointe aux autres modalités dans ce type d'échelle.

Les réponses dichotomiques (oui/non ou vrai/faux ou présence/absence) peuvent être considérées comme faisant partie des échelles nominales, cependant il est fréquent de pouvoir établir une relation d'ordre entre deux modalités, c'est pourquoi elles sont souvent assimilées à des échelles d'intervalles. Par exemple, à la question "je souffre" pour laquelle une réponse dichotomique est proposée en oui/non, une relation d'ordre peut être établie : si je ne souffre pas, je souffre moins que si je souffre !

Remarque :

Le mode (voir chapitre III, paragraphe 2.3.3.) est le seul indice de position que l'on puisse utiliser lors des échelles nominales. Plusieurs modes peuvent cependant être observés pour une même série de données.

2.4.2. Echelles ordinales

Les modalités d'une échelle ordinale **sont liées par une relation d'ordre, sans que l'écart entre ces modalités soit précisé.** Dans ce cas on peut alors classer les différentes modalités les unes par rapport aux autres.

Par exemple, le caractère qualitatif "aimer les statistiques" peut prendre les modalités : pas du tout, très peu, un peu, beaucoup ; pour lesquelles une relation d'ordre peut être proposée ; de même que pour le caractère qualitatif "intensité de la douleur" (nulle, faible, soutenable, insoutenable), "statut économique", "rang de préférence" parmi plusieurs objets à classer (par exemple des items à classer pour établir le poids des items).

L'ordinalité permet de calculer une médiane.

2.4.3. Echelles d'intervalles, numériques ou quantitatives

Les modalités d'une échelle d'intervalle sont liées par une relation d'ordre et permettent toutes opérations arithmétiques, l'écart entre deux modalités ordonnées étant égal. Par exemple, le caractère qualitatif "classe d'âge" peut prendre les modalités : 20 - 24 ans, 25 - 29 ans, 30 - 34 ans, 34 - 39 ans ; pour lesquelles les intervalles sont de taille égale ; de même que pour le caractère quantitatif "taille des sujets" caractérisé par une même unité de mesure (par exemple le mètre), ou tout caractère quantitatif.

Les échelles d'intervalles ne permettent pas toujours de déterminer une médiane.

2.5. Recueil des données

Différents types de recueil de données sont possibles :

- l'auto-questionnaire, où le sujet observé répond lui-même aux questions,
- l'entretien, où c'est l'enquêteur qui note les réponses des sujets observés ou note ses propres observations à propos des sujets.

Différents moyens de recueil sont possibles :

- par téléphone, où la saisie automatique simultanée est envisageable,
- à partir d'entretien, où dans ce cas les conditions doivent être homogènes pour l'ensemble des sujets,
- les auto-questionnaires permettant un recueil automatique, par le sujet lui-même.

3. Traitement des données

Le traitement des données consiste en leur **saisie** sur ordinateur, leur **codage** pour rendre l'analyse intelligible, et la vérification de leur cohérence (comprenant les erreurs de saisie). Cette étape est primordiale et représente généralement la partie la plus longue du travail en statistique. Il est préférable de la compléter par une analyse descriptive univariée des données

comme cela est présenté dans le chapitre IV, afin de retrouver les erreurs de saisie, les données manquantes, etc.

3.1. Saisie des données

La saisie des données consiste à enregistrer les réponses aux questionnaires sur ordinateur. Elle permet de vérifier la qualité de remplissage des questionnaires. **L'étape de la saisie peut être source d'erreurs importantes,** soit que la personne qui fait la saisie se trompe de touche au clavier, soit qu'elle ait mal lu le résultat du questionnaire. Pour éviter les erreurs de saisie, il faut donc déjà que les réponses à mettre sur ordinateur ne prêtent pas à confusion (coder les réponses d'une variable 'O' ou 'P' est risqué puisque les touches sont côte à côte sur le clavier, une réponse positive marquée 'O' peut être confondue avec zéro) et que le module de saisie soit le plus proche possible du questionnaire (si les réponses d'un questionnaire sont rangées identiquement sur plusieurs lignes, le codeur va finir par se tromper de ligne du questionnaire ou se tromper de ligne sur l'ordinateur).

A ce titre, il faut éviter d'utiliser directement des codages conventionnels plutôt que des réponses explicites au moment de remplir un questionnaire. On sait en effet que les statistiques se calculent sur des chiffres, mais il vaut mieux éviter de faire saisir par exemple 1 pour les réponses "oui" et 2 pour les réponses "non" ; tout d'abord parce que le codeur peut se tromper de chiffres, mais aussi parce que ce recodage est facile et très rapide à faire sur ordinateur avec les logiciels de statistiques.

Une confusion est souvent faite entre ***donnée manquante*** et réponse négative, qui gêne ensuite l'exploitation d'une enquête : une donnée manquante est une réponse non fournie dans le questionnaire, elle se code différemment de la réponse négative. On peut établir des codes spécifiques d'une donnée manquante tels que 'I' (Inconnu) contre 'N' pour une réponse négative, ou '-1', '-99' contre '0' pour une réponse négative. Supposons qu'on propose un auto-questionnaire à des psychologues pour chacun de leurs patients, et qu'une des questions soit "votre patient a-t-il consulté un autre psychologue au cours de sa vie ?". Si la réponse n'est pas cochée, on ne sait pas s'il s'agit d'une réponse négative, ou bien si le psychologue n'a pas interrogé le patient à ce sujet (donnée manquante). On pourra alors compter le nombre de patients qui ont consulté différents psychologues, mais on ne pourra pas savoir quelle est la proportion de patients qui en ont consulté au moins un autre (rapport du nombre de réponses positives sur le nombre de réponses négatives et positives) parmi les sujets qui ont fourni une réponse... Le fait de ne pas pouvoir faire cette distinction peut rendre impossible l'exploitation de certaines questions parmi les plus intéressantes.

Le pourcentage de données manquantes doit être surveillé dans le déroulement d'une enquête, car un pourcentage supérieur à 10 % peut être le signe d'une question mal posée (enquête pilote). Les études pilotes ont pour rôle de permettre de reformuler les questions, avant de lancer l'enquête à plus grande échelle. Un fort pourcentage de données manquantes pour une variable donnée peut conduire à la retirer de l'analyse des résultats.

On peut intégrer dans les questionnaires des vérifications de cohérence comme demander l'âge et la date de naissance à deux étapes du questionnaire. Il est recommandé de pouvoir identifier le répondant sur chaque fiche en les numérotant quand une personne fait les mêmes erreurs sur plusieurs questionnaires ou que des réponses sont manquantes.

En général, les données des questionnaires ou des cahiers d'observations sont saisies à la main, mais il existe des modules de saisie sur ordinateur. Ces modules permettent de présenter une page de saisie sur ordinateur qui ressemble au questionnaire papier, contrôlent si la variable saisie fait partie des réponses possibles indiquées préalablement (et, dans le cas contraire, l'ordinateur affiche un message d'erreur et propose à nouveau la même saisie).

La double saisie permet *in fine* de garantir une bonne qualité des données saisies. Il s'agit de faire une deuxième saisie de tous les questionnaires : quand une variable a été codée différemment, on fait encore une saisie jusqu'à obtenir un résultat cohérent.

3.2. Codage et recodage des données

Les données analysées par les logiciels de statistiques se présentent souvent sous la forme de matrices. Après la saisie, cette matrice est généralement un fichier ASCII avec une ligne par sujet observé et les variables présentées en colonnes.

Pour obtenir des statistiques, il faut coder les variables alphanumériques en numériques, soit directement sur le fichier ASCII lu, soit dans le programme du logiciel de statistique utilisé. Le plus couramment utilisé est par exemple la transformation des réponses "oui/non" en "1/0", qui permet ensuite d'obtenir le pourcentage de réponses positives avec l'intervalle de confiance.

Les données qui vont présenter le plus de difficulté sont les **données manquantes et les dates**. Les données manquantes (codées généralement par un point '.' ou -99) doivent être clairement différenciées des réponses négatives (codées généralement 0). En effet, si par exemple on veut connaître le pourcentage de réponses positives lorsque l'on a codé indifféremment 0 les réponses négatives et les non réponses, le résultat obtenu est faux. Il faut compter les réponses positives par rapport aux réponses posi-

tives/négatives en excluant les non réponses (dont on ne sait pas si elles auraient été positives ou négatives). Pour les dates, il y aura à chaque logiciel un format de lecture défini mais il peut varier selon le pays. Par exemple, les données sont généralement lues sous le format MM-JJ-AA (comme 12-01-98 pour le 1er décembre 1998) aux USA et JJ-MM-AA en Europe.

Pour une question dont la réponse négative (codée zéro) a un effet positif, on pourra préférer inverser le codage en positif (coder 1), afin de regrouper des questions qui ont un effet positif. Lors de la constitution d'une échelle de mesure de la qualité de vie, on pourra recoder les réponses afin que l'ensemble de celles-ci codées '1' indiquent une amélioration de la qualité de vie plutôt qu'une diminution. Dans ce cas, le score obtenu sera d'autant plus grand que la qualité de vie sera meilleure (ou *vice versa*).

Pour rendre les données exploitables, le **recodage** permet aussi de dédoubler une variable qui contient deux informations pertinentes. Si on enquête par exemple auprès de psychologues pour savoir combien de patients ils refusent chaque année, on aura ceux qui auront répondu "zéro patients" et ceux qui auront refusé 1,2,..,n patients. Il est intéressant de regrouper ceux-ci dans un seul groupe (réponses positives 'O') pour les opposer à ceux-là (réponse négative 'N') ; on peut alors donner le pourcentage de psychologues qui "refusent des patients dans l'année" comme étant le rapport des 'O'/('O'+'N'). On peut créer une deuxième variable "combien de patients avez-vous refusés cette année" pour ceux qui ont répondu 'O' à la question précédente et en tirer la distribution de fréquence du nombre de psychologues par nombre de patients refusés.

4. Analyse statistique des données

Des événements imprévus peuvent conduire à réadapter le plan d'analyse décrit dans le protocole. Cela dans le cas où certaines données n'ont pu être recueillies ou quand les résultats ne sont pas ceux attendus et demandent des analyses complémentaires, voire des données supplémentaires à analyser.

En dehors de toute difficulté rencontrée, il s'agit à cette étape **d'adapter simplement le plan d'analyse préconçu aux données recueillies**. Chaque test proposé doit alors être réalisé en s'attachant tout particulièrement à vérifier les conditions d'applications. Nous proposons alors, dans le paragraphe 6 suivant, un résumé des procédures statistiques à mettre en place lors de l'analyse des données, dans le cas de tests d'hypothèses.

▩ *Remarque :*

Le risque d'erreur de première espèce est modifié quand l'étude est interrompue ou augmentée.

5. Interprétation des résultats

L'interprétation des résultats des tests statistiques, et plus exactement des conclusions issues des tests statistiques, doit être émise avec beaucoup de **prudence**, en s'attachant tout particulièrement au contexte de l'étude et au protocole. Il est essentiel de rattacher les conclusions apportées au type de sujets étudiés et à la population de référence (si les données sont représentatives), et donc d'**éviter des généralisations et des relations de causalité abusives**.

L'interprétation des résultats ne pourra mettre en évidence un effet de causalité ou une liaison entre les variables étudiées que si "**toutes choses sont égales par ailleurs**". C'est-à-dire que les groupes étudiés doivent être comparables, hormis ce qui concerne la ou les variables étudiées (tous les facteurs doivent être théoriquement contrôlés). On ne pourra généraliser l'effet de causalité observé, à partir d'échantillons, à l'ensemble de la population de référence, que **si et seulement si les échantillons sont représentatifs** et ont donc été obtenus par tirage au sort. C'est-à-dire si toutes choses sont égales par ailleurs, hormis sur la variable dont l'effet est étudié (voir le paragraphe 3 sur l'échantillonnage d'une population du chapitre III). Bien entendu, on ne pourra attribuer un effet de causalité entre A et B que si A précède B et que si B survient seulement en présence de A.

Toute interprétation des résultats dépend de la constitution des échantillons. Par exemple, lorsque l'on conclut à une différence significative entre deux distributions observées du Quotient Intellectuel selon la classe d'âge, on ne pourra conclure que le QI dépend de la classe d'âge en règle générale que si les échantillons observés ont été constitués par tirage au sort et peuvent donc être considérés comme représentatifs de chaque classe d'âge de la population de référence.

6. Procédure résumée des tests d'hypothèses

L'analyse statistique des données observées comprend trois parties majeures :

1) les statistiques descriptives destinées à préparer la véritable tâche du statisticien en visualisant et en synthétisant les renseignements recueillis ;

2) la mise en ordre, le regroupement, puis la réduction des données à des statistiques (moyenne, écart-type, étendue) afin d'établir clairement les hypothèses à tester ;

3) le test des hypothèses afin d'interpréter les résultats en évaluant les limites des conclusions (il s'agit des statistiques inférentielles). Il est

souhaitable, lors de cette étape, de valider ces conclusions par exemple par les résultats obtenus dans la littérature.

▩ *Difficulté et danger des statistiques :*

Mesure-t-on bien ce que l'on veut mesurer ? Le concept à mesurer n'est-il pas dévié, déformé, transformé, biaisé par des facteurs de confusion ? C'est pourquoi il est nécessaire d'être attentif à la constitution des échantillons et à l'élaboration du cahier d'observation qui détermine les concepts à mesurer.

Les dangers résident également dans l'interprétation des résultats qui nécessite beaucoup de prudence pour toute généralisation. Il faut se méfier des conclusions hâtives.

■ *Procédure des tests statistiques :*

Afin de décrire chaque étape à mettre en œuvre dans la réalisation des tests d'hypothèses, **nous présentons ici une procédure générale des tests en 12 points :**

1. Déterminer la nature des variables à étudier (qualitatives ou quantitatives). Dans le cas de variables qualitatives, déterminer s'il existe une relation d'ordre.

2. Déterminer si les données sont appariées ou non, et s'il s'agit d'une comparaison de deux ou plus de deux groupes observés, ou d'un groupe observé à une population théorique.

3. Poser les hypothèses nulle et alternative, H_0 et H_1. Concernant l'hypothèse alternative, déterminer si l'on procède à un test bilatéral ou à un test unilatéral (à droite ou à gauche).

4. Déterminer le test adapté aux hypothèses et au type de données.

5. Vérifier les conditions d'application du test approprié, et si celles-ci ne sont pas vérifiées, procéder éventuellement à un test non paramétrique adapté aux données.

6. Fixer le seuil de signification de la différence (le plus souvent : $\alpha = 5\ \%$), c'est-à-dire le risque d'erreur de première espèce, ou encore la probabilité de rejet de l'hypothèse nulle H_0 à tort.

7. Déterminer le paramètre à calculer.

8. Déterminer le paramètre théorique, lu dans la table statistique correspondant au test, selon le seuil d'erreur α choisi.

9. Déterminer où se situe le paramètre calculé par rapport au paramètre théorique (inférieur ou supérieur) et conclure au rejet ou non de l'hypothèse nulle. Si on rejette l'hypothèse H_0, alors la différence ou l'inégalité est significative (S). Sinon on conclut que la différence est non significative (NS).

10. Si on conclut à une différence ou une inégalité significative, rechercher α', le degré de signification de cette différence, c'est-à-dire la probabilité p jusqu'à laquelle on conclut encore à une différence significative (notation : $p < \alpha'$).
11. Interpréter les résultats en fonction des hypothèses et de la constitution du ou des échantillons.
12. Déterminer les analyses statistiques complémentaires en fonction des résultats. Par exemple, si on conclut à une différence significative entre plusieurs moyennes (plus de 2), rechercher où se situe cette différence entre les moyennes deux à deux.

7. Les logiciels statistiques

Lorsqu'on est amené à manipuler un nombre important de données, et surtout lorsque l'on doit réaliser plusieurs analyses statistiques (tests) sur un même jeu de données, il est préférable d'utiliser un logiciel statistique, afin de gagner du temps.

Il y a encore peu de temps, la réalisation des calculs statistiques était confiée à un informaticien sur d'imposants ordinateurs. Ces "gros travaux" statistiques sont maintenant accessibles à nos lecteurs sur un petit ordinateur PC-compatible ou un Macintosh. Ce phénomène ne peut qu'aller en s'accentuant et chacun peut avoir à sa disposition des outils statistiques de plus en plus performants et conviviaux.

Comme ces logiciels ne proposent pas encore une assistance efficace à l'utilisateur (avec des programmes d'intelligence artificielle), le risque est que le logiciel fonctionne souvent comme une "boîte noire" : vous choisissez deux colonnes de données, un test dans le menu et obtenez un résultat, sans pouvoir apprécier si le test est adéquat, ou la méthode de calcul du logiciel appropriée.

Nous vous conseillerons donc de ne pas céder à la facilité, mais de **choisir le test adapté** à l'aide de ce livre (voir l'arbre décisionnel du chapitre I), puis de retrouver la marche à suivre sur l'ordinateur (ou sur les listings). N'hésitez pas à passer du temps sur la documentation du logiciel pour comprendre ce que fait la "boîte noire" quand vous choisissez un test dans le menu. On voit souvent celui qui découvre un logiciel le faire tourner avec ses données jusqu'à obtenir des résultats qui lui paraissent convenables. Cela peut conduire à un mauvais choix, n'oubliez pas que le statisticien classique est un parieur, qui fixe ses hypothèses *a priori*, et non en fonction des résultats que lui fournit l'ordinateur.

Dans les logiciels que nous vous présentons, on peut distinguer les logiciels professionnels, complets ou spécialisés et les logiciels personnels. Les logiciels professionnels SAS ou BMDP ont été créés pour gros ordinateurs, mais peuvent être adaptés sur ordinateur personnel moyennant un investissement de 8 000 à 15 000 F (généralement sous forme d'abonnement avec un tarif dégressif d'année en année). Avec un investissement plus raisonnable, on peut trouver plusieurs logiciels permettant de réaliser l'ensemble des tests décrits dans ce livre, tels qu'EpiInfo (gratuit), S-plus, Statview, Systat, Statbox (+Excel), SPSS. Certains logiciels sont spécialisés : GLIM pour les modèles linéaires (analyse de variance, régression), Extend ou Alice (aide à la décision), Sphinx ou Question (enquêtes, analyses multivariées).

Nous citons ci-dessous ces logiciels selon la gamme de prix : A *(plus de 8 000F)*, B *(de 4 000 à 8 000F)*, C *(moins de 4 000F)*. Il ne nous est pas possible de vous faire une liste exhaustive et une critique des logiciels statistiques, mais nous espérons vous aider pour votre choix. Nous vous convions à demander quels sont les logiciels de statistique que propose le vendeur de votre matériel informatique et à vous renseigner aussi auprès de votre entourage.

7.1. Logiciels professionnels

Les logiciels professionnels ont été développés sur de gros systèmes informatiques comme Unix, VMS jusqu'à des modules élaborés de traitement des données, voire de saisie ; ils permettent de réaliser toutes les statistiques jusqu'à l'analyse de séries temporelles, les statistiques spatiales, etc. Ils peuvent être utilisés en même temps que des langages comme C, pascal ou fortran. Le plus connu est SAS ; BMDP a été racheté par SPSS Inc. (qui propose un autre logiciel de statistique appelé SPSS) ; S-plus est un langage élaboré et convivial comme l'est Mathematica pour les mathématiques ; NAG est une librairie de fonctions mathématiques qui propose deux logiciels de statistiques. Ces logiciels tournent aussi sur PC-compatibles dans des versions moins sophistiquées.

SAS : c'est un logiciel complet de statistiques que l'on peut acheter par modules (tarif A) et aussi un système de gestion de base de données. Il existe en version PC-compatible ou une version simplifiée JMP pour Macintosh (tarif B).

SAS Institute - Domaine de Grégy - Boîte postale 5 - 77166 Grégy-sur-Yerres (France)

STATA : On peut aussi citer STATA qui existe pour Unix, DOS, Windows et Mac et qui remplit les mêmes fonctions que SAS (tarif B).

Renseignements : www.ritme.com

BMDP New System : presque aussi complet mais sans gestion de données. Editions professionnelle (tarif A) ou personnelle (tarif C).

SPSS France SAR - 72-74, avenue Edouard-Vaillant - 92100 Boulogne (France)

S-plus : langage de troisième génération très complet (tarif A), utilisé couramment pour les statistiques bayésiennes. Il peut être utilisé sur PC-compatible (tarif A).

StatSci Europe- distributor of S-PLUS - 52 Sandfield Road - Headington - OXFORD OX3 7RJ (England)

Librairie NAG (logiciels GLIM, GENSTAT)

GENSTAT 3.2 : Statistiques de base, régression, analyse multivariée, analyse de survie, séries temporelles, géostatistiques.

GLIM 4.0 : logiciel interactif d'investigation pour l'analyse de régression (en particulier poissonnienne), anova, modèles de survie (tarif C).

NAG - Espace III - 62, boulevard Frederic-Arnaud - 09200 Saint-Girons - Toulouse (France)

7.2. Logiciels sur ordinateur personnel

Ces logiciels sont largement suffisants pour l'analyse statistique courante de vos données. Nous vous conseillons cependant d'utiliser des versions PC-compatibles des logiciels professionnels décrits ci-dessus, afin d'être familiarisé avec eux si vous êtes amené à les utiliser par la suite.

Certains de ces logiciels existent en version "Student", moins coûteuse.

SPSS 6.1 (Macintosh, PC) : c'est un logiciel complet et convivial pour les non statisticiens (tarif B).

SPSS France SAR - 72-74, avenue Edouard-Vaillant - 92100 Boulogne (France)

SYSTAT 5.2 (Macintosh, PC) : a été racheté par SPSS Inc. (tarif B).

STATVIEW 4.5 (Macintosh, PC) : logiciel convivial pour les non statisticiens (tarif B).
Abacus Concepts, Inc. - 1918 Bonita Avenue - Berkeley, CA 94704-1014 (USA)

MINITAB (Macintosh) : (tarif C).
Minitab Ltd. - 3 Mercia Business Village - Torwood Close - Westwood Business Park - CV4 8HX Coventry (ENGLAND)

NCSS (PC) : (tarif B).
NCSS Statistical Software - 329 North 1000 East - Kaysville, Utah 84037 (USA)

Statistica (PC) : Logiciel statistique convivial, mais ne permet pas d'ajouter des individus supplémentaires lors de l'analyse multiple (tarif C).

EXCEL 5.0 (Macintosh, PC) : c'est un tableur avec lequel vous pouvez programmer vos calculs de corrélation, de chi-deux, etc (tarif C). Ce n'est pas un logiciel de statistique mais il contient des fonctions statistiques ; il peut servir à résoudre les exercices du livre et faire des représentations graphiques qui vous aideront à mieux comprendre le cours (vous pouvez par exemple programmer une corrélation entre deux variables quantitatives, calculer le coefficient de corrélation et le test t). Vous pouvez y ajouter un module statistique STATBOX.

7.3. Logiciels gratuits

Les logiciels gratuits peuvent être excellents et couramment utilisés, c'est le cas du seul que nous connaissons, EpiInfo, développé par le CDC d'Atlanta et l'OMS (Organisation Mondiale de la Santé). La contrepartie de leur coût nul est souvent le manque de documentation, l'existence de bugs...

Var-3 (PC) : Logiciel d'analyse de variance simple.
Institut de Psychologie Henri Pieron - 28, rue Serpente - 75006 Paris

ADAD (PC) : Logiciel d'analyse statistique remarquable pour l'analyse multivariée.

EpiInfo 6.0 (PC) : logiciel d'enquête en épidémiologie avec bon module de saisie et statistiques univariées. Il peut être récupéré sur les sites ftp suivants :
CDC d'Atlanta (USA) : http://www.cdc.gov/

Serveur EpiWeb : http://www.univ-lille2.fr/epiweb/epif/.

ADE-4 4.0 (Macintosh) :
Laboratoire de biométrie, génétique et biologie des populations
Université Claude Bernard - Lyon I - 43, bd du 11 novembre 1918 - 69622 Villeurbanne cedex (France)

MacAnova 4.0 (Macintosh) : logiciel d'analyse statistique interactif
Gary Oelhert et Christopher Bingham - University of Minnesota, School of statistics - 352 Classroom-Office building - 1994 Buford Avenue, St. Paul - MN 55108 (USA)

ViSta 4.2 (Macintosh) :
Forrest Young - University of North Carolina - Department of psychology - CB 3270, Davie Hall, Chapel Hill - NC 27599-3270 (USA)

7.4. Sites Internet

Nous vous signalons aussi l'existence de nombreux serveurs statistiques sur Internet. Dans la mesure où ils peuvent changer d'adresse ou disparaître, nous vous conseillons de rechercher ces adresses par mots-clés avec des serveurs comme AltaVista (http://www.altavista.telia.com/) sur des critères multiples comme "statistical software" (plus de 5 000 réponses en 1997), AND "archive" (environ 1 000 réponses que vous pouvez consulter avant de restreindre par un mot-clé supplémentaire). Vous trouverez aussi environ 1 000 réponses si vous cherchez en France : "logiciels and statistiques and france", et moins de 100 si vous ajoutez "and psychologie".

Voici quelques exemples de ce que l'on peut obtenir sur Internet :

— des listes de discussion sur les statistiques comme par exemple STAT-L :
The WWW Virtual Library (http://www.stat.ufl.edu/vlib/statistics.html),
L'index Statlib (http://teseo.unipg.it/statlib/),
Statistics on the Web (http://www.execpc.com/~helberg/statistics.html),
Statlinks (http://151.121.66.126/sigstat/statlink.htm#link0)

— des programmes de cours de statistiques tels que :
La licence de psychologie de Rouen
(http://epeire.univ-rouen.fr/diplomes/lic-psy.htm),
L'université de Strasbourg
(http://www-ulp.u-strasbg.fr/actualites/bulletin_vert/bv914.html),

Le département de psychologie de l'université d'York au Canada
(http://www.yorku.ca/dept/psych/lab/psy3030/),
L'université d'Arizona aux USA
(http://olam.ed.asu.edu/~glass/502/home.html),
L'université de Melbourne en Australie
(http://www.stats.mu.oz.au:8001/discday/discday.html),
Hyperstat Contents
(http://www.ruf.rice.edu/~lane/hyperstat/contents.html).

— *des sites pour logiciels comme :*
Journal of Statistical Software
(http://www.stat.ucla.edu/journals/jss/JSS/0053.html),
Software Packages (http://www.stats.umanitoba.ca/software.html),
A Guide to Statistical Computing Resources on the Internet
(http://asa.ugl.lib.umich.edu/chdocs/statistics/stat_guide_home.html).

— *des dictionnaires ou des exemples comme :*
Statistics Every Writer Should Know (http://nilesonline.com/stats//),
Data and Story Library (http://lib.stat.cmu.edu/DASL/),
Chance database
(http://www.geom.umn.edu/docs/snell/chance/welcome.html),
StatLearn (http://www.stat.psu.edu/statlearn/),
Statlinks (http://151.121.66.126/sigstat/stated.htm).

CHAPITRE III

Des probabilités vers les statistiques

1. Introduction aux probabilités 63
1.1. Calcul combinatoire 63
1.1.1. Permutations 63
1.1.2. Arrangements 64
1.1.3. Combinaisons 65
1.2. Calcul de probabilités 67
1.2.1. Théorie des probabilités 67
1.2.2. Notations 68
1.2.3. Axiomes du calcul des probabilités 69
1.3. Probabilités conditionnelles 71
1.4. Indépendance 73
1.5. Arbre des probabilités 74
1.6. Théorème de Bayes 76

2. Introduction aux statistiques 79
2.1. Variable aléatoire, fonction de répartition 79
2.2. Densité de probabilité d'une variable aléatoire continue 81
2.3. Statistiques de position : Moyenne, Médiane, Mode 81
2.3.1. La moyenne 82
2.3.2. La médiane 83
2.3.3. Le mode 84
2.4. Statistiques de dispersion : Etendue, Ecart-moyen, Ecart-type, Variance, Covariance 86
2.4.1. L'étendue 86
2.4.2. L'écart-moyen 86
2.4.3. L'écart-type 87
2.4.4. La variance 88
2.4.5. Le cœfficient de variation 90
2.4.6. La covariance 90
2.4.7. La corrélation 91

3. L'échantillonnage d'une population *91*
3.1. Randomisation *92*
3.2. Echantillons appariés *93*
3.3. Cas-témoin *93*

4. Lois de probabilité *93*
4.1. Loi binomiale *93*
4.2. Loi de Poisson *94*
4.3. Loi normale *95*
4.3.1. Formulation *95*
4.3.2. Utilisation de la table de la loi normale N(0,1) *98*
4.3.3. Applications *101*

5. Notion d'intervalle de confiance *103*
5.1. Variance d'une moyenne *104*
5.2. Intervalle de pari d'une moyenne *106*
5.3. Intervalle de confiance d'une moyenne *107*
5.3.1. Cas de grands échantillons, $n \geq 30$ *107*
5.3.2 Cas de petits échantillons $n < 30$ *108*
5.4. Intervalle de pari d'un pourcentage *108*
5.5. Intervalle de confiance d'un pourcentage *109*

6. Exercices *110*
6.1. Exercices corrigés *110*
6.2. Exercices non corrigés *111*
6.3. Solutions *113*

CHAPITRE III

Des probabilités vers les statistiques

1. Introduction aux probabilités

Pourquoi consacrer un paragraphe au calcul des probabilités ? Parce que **c'est sur la base des calculs de probabilités que s'est développée la Statistique** et qu'il faut acquérir quelques notions de base. Vous aurez besoin de ces notions pour comprendre les bases théoriques de ce cours, plus que pour appliquer les tests statistiques : si cela vous paraît rébarbatif, pensez qu'il peut être amusant d'appliquer le calcul des probabilités pour calculer ses chances au loto, ou mieux jouer aux cartes !... Est-ce qu'au poker, j'ai plus de chance d'obtenir un poker de valet ou une suite à trèfle, avec une main de 9, 10 et valet de trèfle et deux autres valets ?

1.1. Calcul combinatoire

Nous verrons dans le paragraphe 1.2 qu'**une probabilité peut s'écrire comme le nombre de cas favorables rapporté au nombre de cas possibles**. On sait par exemple qu'il existe une chance sur 6 d'avoir un 4 au lancer de dé, mais quelles sont nos chances avec 3 dés, avec 10 dés ? Le calcul combinatoire va nous permettre notamment de déterminer les nombres de cas possibles et les nombres de cas favorables.

1.1.1. Permutations

Soit n objets a_1, a_2, ..., a_n. On peut les ranger de P_n façons. **P_n est le nombre de permutations de n objets.**

Par exemple, P_5 est le nombre de façons de ranger 5 objets différents dans 5 cases ou 5 tiroirs :

Pour ranger 5 objets dans 5 cases, on dispose de 5 possibilités de rangement pour le premier objet, puis une fois le premier rangé, on dispose de 4 possibilités de rangement pour le deuxième objet, puis on dispose de 3 possibilités de rangement pour le troisième objet, puis 2 pour le quatrième, et enfin on ne dispose plus que d'une possibilité de rangement pour le cinquième et dernier objet. Soit : 5 x 4 x 3 x 2 x 1.

Ainsi, $\boxed{P_n = n \times (n-1) \times \ldots \times 3 \times 2 \times 1 = n!}$

et par convention $P_0 = 0! = 1$

Notons que n ! se lit factorielle n, 0 ! se lit factorielle 0.

• Exemple 1 :

Quel est le nombre de permutations possibles des lettres du mot "PSYCHÉ" ?
Les 6 lettres du mot "PSYCHÉ" peuvent être rangées ou disposées de P_6 façons, c'est-à-dire :
6 ! = 6 x 5 x 4 x 3 x 2 x 1 = 720 façons.

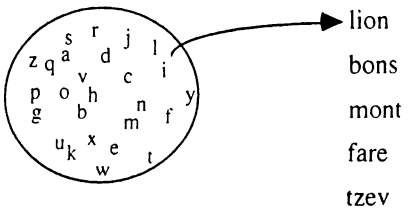
• Exemple 2 :

Lors d'une course de chevaux, où 10 chevaux sont au départ, quel est le nombre d'ordres d'arrivées possibles, en supposant qu'il ne peut y avoir d'ex aequo ?
Il existe 10 ! ordres d'arrivées possibles, soit 3 628 800 rangements possibles.

1.1.2. Arrangements

Soit n objets a_1, a_2, ..., a_n. Le nombre de façons de tirer p objets parmi ces n objets, **en tenant compte de l'ordre dans lequel on les tire, est le nombre d'arrangements de p objets parmi n, que l'on note** A_n^p.

La figure suivante visualise 5 exemples de tirage de 4 lettres parmi les 26 lettres de l'alphabet. Le nombre de façons de tirer 4 lettres parmi 26, en tenant compte de l'ordre dans lequel on les tire, est alors 26 possibilités pour la première lettre, puis 25 pour la seconde, puis 24 pour la troisième, et enfin 23 pour la quatrième. Soit $A_{26}^4 = 26 \times 25 \times 24 \times 23$.



Alors, $\boxed{A_n^p = n\,(n-1)\,(n-2)\ldots(n-p+1) = \dfrac{n!}{(n-p)!}}$ avec $p \leq n$

notons que : $A_n^n = n!$, $A_n^1 = n$, $A_n^0 = 1$.

• Exemple 1 :

Lors de la sélection de 5 candidats classés parmi 20 candidats se présentants à un concours, quel est le nombre de sélections possibles (en supposant qu'il ne peut y avoir d'ex aequo) ?
Il existe $A_{20}^5 = 20 \times 19 \times 18 \times 17 \times 16$ sélections possibles, soit 1 860 480.

• Exemple 2 :

Lors d'un tiercé, où 15 chevaux sont en course, quel est le nombre de tiercés gagnants possibles, en supposant qu'il ne peut y avoir d'ex aequo ? Même question pour le nombre de quintés possibles ?
Il existe $A_{15}^3 = 15 \times 14 \times 13$ tiercés possibles, soit 2 730.
Pour le quinté, il y a 12 x 11=132 fois plus d'arrangements (en tout 360 360 possibilités) : commencez donc par vous exercer au tiercé !

1.1.3. Combinaisons

Soient n objets $a_1, a_2, \ldots, a_n$. Le nombre de façons de tirer p objets parmi ces n objets, **sans tenir compte de l'ordre dans lequel on les tire, est le nombre de combinaisons de p objets parmi n, que l'on note** C_n^p ou $\binom{n}{p}$.

$$\boxed{C_n^p = \frac{A_n^p}{p!} = \frac{n!}{(n-p)!\,p!}}$$

$$\boxed{C_n^p = \frac{n\,(n-1)\,(n-2)\ldots(n-p+1)}{p!}} \quad \text{avec } p \leq n$$

notons que : $C_n^n = 1$, $C_n^1 = n$, $C_n^0 = 1$.

Ici p ! correspond au nombre de permutations de p objets, c'est-à-dire au nombre de façons de ranger ou disposer p objets (ou d'ordonner p objets).

■ ***Propriété :***

A chaque fois que l'on prend p objets, on en laisse n-p dans l'ensemble initial. Le nombre de façons de prendre p objets parmi n objets au total, C_n^p, est donc identique au nombre de façons de laisser n-p objets parmi n, alors $C_n^p = C_n^{n-p}$.

■ **Pour bien distinguer une combinaison d'un arrangement**, on peut considérer qu'une **combinaison** est le résultat d'un tirage de p éléments *simultanément* parmi n, alors que pour un **arrangement**, on tire les p éléments l'un après l'autre *en s'intéressant à l'ordre* dans lequel on les obtient. Nous avons vu qu'il existe 2 730 tiercés possibles (3 chevaux parmi 15) en tenant compte de l'ordre d'arrivée, mais il n'existe que 455 (2730/3 !) tiercés possibles sans tenir compte de l'ordre. Il est bien plus facile de déterminer la bonne arrivée de 3 chevaux dans le désordre ! En termes de probabilités (que nous abordons dans les pages suivantes), nous pouvons dire qu'il est plus probable de gagner au tiercé sans trouver l'ordre d'arrivée, que de le gagner dans l'ordre.

• Exemple 1 :

Quel est le nombre de façons de tirer 3 numéros parmi 10 ?

$$C_{10}^3 = \frac{10 \times 9 \times 8}{3!} = \frac{10 \times 9 \times 8}{3 \times 2} = 10 \times 3 \times 4 = 120$$

• Exemple 2 :

Combien d'échantillons différents peut-on constituer en tirant au hasard 40 sujets parmi 100 sujets ?

$$C_{100}^{40} = \frac{100 \times 99 \times 98 \times \ldots \times 61}{40!}$$

$$C_{100}^{40} = \frac{100!}{60!\,40!} = \frac{100 \times 99 \times 98 \times \ldots \times 61}{40!} = \frac{100!/\,60!}{40!} = 1{,}37462 \cdot 10^{28}$$

1.2. Calcul de probabilités

La théorie des probabilités permet de modéliser les phénomènes liés au hasard (phénomènes aléatoires, stochastiques, non déterministes) et de prédire la survenue des événements en terme de fréquence d'apparition.

1.2.1. Théorie des probabilités

Connaissant le modèle probabiliste d'un jeu de hasard comme un lancer de pièce (jeu de pile ou face), on peut prédire les fréquences d'apparition d'événements comme le nombre de fois où l'on obtient pile sur 2 lancers. Ou encore, connaissant le modèle probabiliste des jeux de cartes tels que la belote (qui se joue avec un jeu de 32 cartes), on peut prédire les fréquences d'apparition d'événements comme : obtenir un carré de valet dans une main de huit cartes, une fois toutes les cartes distribuées.

■ *Les échantillons*

Dans ces situations simples, on peut compter exactement le nombre des événements et les énumérer. En revanche, lorsque l'on s'intéresse à des problèmes où de grandes populations sont en cause, il est impossible d'avoir la même approche. Bien que, théoriquement, on puisse étudier une population, pratiquement, on est obligé d'étudier des échantillons de celle-ci, et l'on ne dispose donc pas de toute l'information présente dans la population. L'étude des différences entre des populations (selon les variables considérées), à partir d'échantillons, impose alors de donner *a priori* un modèle probabiliste qui permet de faire "comme si" on avait étudié la population entière. Cette description essentielle de l'intérêt des probabilités en statistique sera plus claire dans l'exemple suivant. Un institut de sondages enquête sur la prévalence des états dépressifs (la prévalence correspondant au nombre de cas à un temps donné). La "vraie" solution serait de demander à toute la population si elle souffre d'une dépression. Mais, la plupart du temps, cela n'est pas pratiquement faisable. On va donc avoir recours à un échantillon de population : par exemple 1 000 individus, et à un "modèle probabiliste" qui va décrire la façon de sélectionner ces individus pour que les résultats obtenus soient représentatifs de la population, ainsi que pour donner une idée de la variation des résultats si on avait sélectionné 1 000 autres patients.

■ *La loi normale*

Le modèle probabiliste le plus utilisé, sur lequel nous reviendrons plus tard, est celui de la ***loi normale ou de Laplace-Gauss***. Si l'on considère qu'une variable étudiée est distribuée normalement dans la population, on pourra alors calculer les propriétés de l'échantillon observé.

■ *Définitions*

Dans le calcul des probabilités, différents termes doivent être définis :

— **Une épreuve** correspond au protocole d'une expérience (c'est-à-dire à la définition des conditions d'une expérience) dont le résultat est aléatoire.

— L'ensemble de tous les résultats possibles d'une épreuve est appelé **ensemble fondamental, noté F.**

— Chaque résultat possible de l'épreuve, c'est-à-dire chaque élément de F, est appelé **événement élémentaire**. Ainsi, l'ensemble F est l'ensemble de tous les événements élémentaires.

— Un **événement** E est un sous-ensemble de F et est composé d'événements élémentaires, de leur union, et/ou de leur intersection.

1.2.2. Notations :

On note l'ensemble de tous les événements : Ω, avec $F \subset \Omega$ (c'est-à-dire F inclus dans Ω).

L'ensemble Ω comprend :
- tous les événements élémentaires,
- leurs complémentaires,
- tous les événements impossibles,
- l'union finie des événements élémentaires.

Soit E_1 et E_2 deux éléments de Ω :

– E_1 élément de Ω, est l'**événement associé** à l'épreuve.

– **L'ensemble vide**, noté $\emptyset$, est appelé *événement impossible* ; et son complémentaire, l'ensemble Ω, est appelé *événement certain* (il correspond à l'événement contraire de $\emptyset$).

– **L'événement contraire** de l'événement E_1 est noté $\overline{E}_1$ en langage ensembliste.

– **La réunion des deux événements** E_1 et E_2, notée $E_1 \cup E_2$ en langage ensembliste, est notée l'événement "E_1 ou E_2" en langage probabiliste, ce qui correspond à l'événement où l'un des deux événements E_1 ou E_2 se produit, ou les deux se produisent à la fois. L'union se traduit par un signe + dans le calcul des probabilités (addition), sous certaines conditions.

– **L'intersection des deux événements E_1 et E_2**, notée $E_1 \cap E_2$ en langage ensembliste, est notée l'événement "E_1 et E_2" en langage probabiliste, ce qui correspond à l'événement où les deux événements E_1 et E_2 se produisent à la fois. L'intersection se traduit par un signe X dans le calcul des probabilités (multiplication), sous certaines conditions, si les événements sont *indépendants*.

– **Si l'intersection entre E_1 et E_2 est vide**, en langage probabiliste, on dit que ces deux événements sont *incompatibles*, c'est-à-dire que si l'un se produit, l'autre ne le peut pas. Dans ce cas, E_1 et E_2 ne sont pas *indépendants*.

• Exemple :

Pour l'épreuve du lancer d'un dé (non pipé) où l'on observe le résultat obtenu ; l'ensemble fondamental est formé par les six résultats possibles d'un tirage : $F = \{1, 2, 3, 4, 5, 6\}$. L'événement correspondant à l'apparition d'un nombre pair est $A = \{2, 4, 6\}$ avec A inclu dans l'ensemble F (que l'on note $A \subset F$). L'événement correspondant à l'apparition d'un chiffre supérieur ou égal à 3 est $B = \{3, 4, 5, 6\}$. L'événement correspondant à l'apparition d'un 6 est $C = \{6\}$. Tous les événements élémentaires sont incompatibles.

$$\Omega = \{\{1\}, \{2\}, \{3\}, \{4\}, \{5\}, \{6\}, \{1,2\}, ..., \{1,2,3,4,5,6\}, \varnothing\}$$

1.2.3. Axiomes du calcul des probabilités :

La probabilité est une fonction qui associe à chaque événement un nombre réel compris entre 0 et 1 et vérifiant :

$P(\Omega) = 1$

$P(E_1 \cup E_2 \cup ... \cup E_n ....) = P(E_1) + P(E_2) + ... + P(E_n) + ...$

si les événements E_1, E_2, ...E_n,..., en nombre éventuellement infini, sont disjoints deux à deux (incompatibles), c'est-à-dire si $E_i \cap E_j = \varnothing$ quels que soient i et j.

Lorsque l'on peut compter les événements élémentaires, une probabilité correspond au rapport suivant :

$$\text{Probabilité} = \frac{\text{Nombre de cas favorables}}{\text{Nombre de cas possibles}}$$

■ **Toute probabilité est comprise entre 0 et 1 :**

– **elle est toujours positive ou nulle** (≥ 0) car elle est le rapport de deux nombres positifs,

– **elle est inférieure ou égale à 1** (≤ 1) puisque le nombre de cas possibles est supérieur ou égal au nombre de cas favorables,

– **la probabilité de l'événement impossible est nulle** ($P(\varnothing) = 0$) puisque dans ce cas le nombre de cas favorables est nul,

— **la probabilité de l'événement certain est 1** (soit $P(\Omega) = 1$) car dans ce cas tous les cas possibles sont favorables.

■ **La probabilité qu'un événement E_1 se produise correspond à 1 moins la probabilité de l'événement contraire $\overline{E}_1$:**

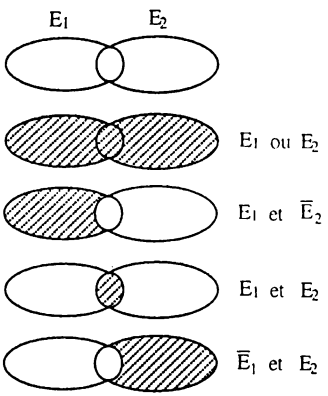
$$P(\overline{E}_1) = 1 - P(E_1)$$

car $P(E_1 \cup \overline{E}_1) = P(\Omega) = 1 = P(E_1) + P(\overline{E}_1)$

Si on note $p = P(E_1)$, alors $P(\overline{E}_1)$ est noté q, donc $q = 1 - p$.

Quels que soient E_1 et E_2 on a : $P(E_1 \text{ ou } E_2) = P(E_1) + P(E_2) - P(E_1 \text{ et } E_2)$

car $E_1 \text{ ou } E_2 = (E_1 \text{ et } \overline{E}_2) \text{ ou } (E_1 \text{ et } E_2) \text{ ou } (\overline{E}_1 \text{ et } E_2)$



or $P(E_1) = P(E_1 \text{ et } E_2) + P(E_1 \text{ et } \overline{E}_2)$

soit $P(E_1 \text{ et } \overline{E}_2) = P(E_1) - P(E_1 \text{ et } E_2)$

et $P(\overline{E}_1 \text{ et } E_2) = P(E_2) - P(E_1 \text{ et } E_2)$

ainsi $P(E_1 \text{ ou } E_2) = P(E_1 \text{ et } E_2) + [P(E_1) - P(E_1 \text{ et } E_2)] + [P(E_2) - P(E_1 \text{ et } E_2)]$

soit $P(E_1 \text{ ou } E_2) = P(E_1) + P(E_2) - P(E_1 \text{ et } E_2)$

On retrouve bien la formule de départ.

Le plus souvent, l'ensemble fondamental est constitué d'un nombre fini N d'événements élémentaires, et dans le cas où tous les événements

élémentaires sont "équiprobables", c'est-à-dire quand ils ont tous la même chance d'être réalisés, alors chaque événement élémentaire a la probabilité $\frac{1}{N}$ de se réaliser.

Aussi pour tout événement E (sous-ensemble d'événements élémentaires de l'épreuve) on a :

$$P(E) = \frac{\text{Nombre de cas favorables}}{\text{Nombre de cas possibles}}$$

soit $$P(E) = \frac{\text{Nombre de cas où l'événement E est réalisé}}{\text{Nombre de résultats de l'épreuve}}$$

ou encore

$$P(E) = \frac{\text{Nombre d'événements élémentaires correspondant à l'événement}}{\text{Nombre total d'événements élémentaires de l'épreuve}}$$

$$P(E) = \frac{\text{Nombre d'éléments du sous-ensemble associé à l'événement}}{\text{Nombre total d'éléments de l'ensemble fondamental}}$$

• Exemple :

Lors de l'épreuve du lancer d'un dé, la probabilité d'obtenir un nombre pair est

$$P(E_1) = \frac{3}{6} = \frac{1}{2} = 0,5$$

avec :

Epreuve : lancer d'un dé et lecture du résultat

E_1 : obtenir un nombre pair, soit $E_1 = \{\{2\} \cup \{4\} \cup \{6\}\}$

$F = \{\{1\}, \{2\}, \{3\}, \{4\}, \{5\}, \{6\}\}$

Et la probabilité d'obtenir au moins un 3 est : $P(E_2) = \frac{4}{6} = \frac{2}{3} = 0,67$

E_2 : obtenir un nombre supérieur ou égal à 3,

soit $E_2 = \{\{3\} \cup \{4\} \cup \{5\} \cup \{6\}\}$

1.3. Probabilités conditionnelles

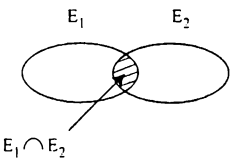
Le calcul des probabilités conditionnelles se base sur les mêmes principes que ceux établis jusqu'ici mais en sachant que certains événements sont déjà réalisés, c'est-à-dire en tenant compte d'**informations déjà connues** dans les calculs. Cela limite alors le nombre de cas possibles. On

conçoit par exemple que si l'on a déjà tiré un roi dans un jeu de cartes, la probabilité d'en tirer un à nouveau est plus faible.

Soit E_1 et E_2 deux événements de l'ensemble fondamental ($E_1, E_2 \in F$) ; on parle de la probabilité de E_1 conditionnellement à E_2, ou encore de **la probabilité de E_1 sachant que E_2 s'est produit**, quand on calcule la probabilité que E_1 se produise alors que l'on sait que E_2 s'est déjà produit. On note cette probabilité $P(E_1/E_2)$ qui correspond par définition à :

Probabilité de E_1 sachant que E_2 s'est produit :

$$P(E_1 / E_2) = \frac{P(E_1 \cap E_2)}{P(E_2)}$$



ou encore

$$P(E_1 / E_2) = \frac{P(E_1 \text{ et } E_2)}{P(E_2)}$$

$$P(E_1 / E_2) = \frac{\text{Nombre d'événements élémentaires de } E_1 \text{ et } E_2 \text{ à la fois}}{\text{Nombre d'événements élémentaires de } E_2}$$

Soit A et B deux événements, on a :

$$P(A / B) = \frac{P(A \cap B)}{P(B)} \quad \text{soit} \quad P(A \cap B) = P(A / B)\, P(B)$$

et $$P(B / A) = \frac{P(A \cap B)}{P(A)} \quad \text{soit} \quad P(A \cap B) = P(B / A)\, P(A)$$

alors $P(A / B)\, P(B) = P(B / A)\, P(A)$

soit $$P(A / B) = \frac{P(B / A)\, P(A)}{P(B)}$$

• Exemple 1 :

Lors d'un lancer de dé, quelle est la probabilité d'avoir obtenu un 2 sachant que l'on a obtenu un chiffre pair ?

$$P(\text{chiffre 2 / chiffre pair}) = \frac{P(\text{chiffre } \{2\} \text{ et chiffre pair})}{P(\text{chiffre pair})} = \frac{P(\text{chiffre } \{2\})}{P(\text{chiffre pair})} = \frac{1}{3}$$

• Exemple 2 :

Quelle est la probabilité de tirer un valet dans un jeu de 52 cartes, sachant que la carte tirée est un cœur ?

$$P(valet / cœur) = \frac{P(valet\ et\ cœur)}{P(cœur)} = \frac{\frac{1}{52}}{\frac{13}{52}} = \frac{1}{13}$$

1.4. Indépendance

Les deux événements E_1 et E_2 sont dits indépendants si la réalisation de l'événement E_1 n'apporte pas d'information sur la réalisation de l'événement E_2, et réciproquement :

	$P(E_1 \text{ et } E_2) = P(E_1) \times P(E_2)$
alors	$P(E_1 / E_2) = P(E_1)$
et	$P(E_2 / E_1) = P(E_2)$

Ainsi, la connaissance de la survenue d'un des deux événements n'apporte aucune information sur la survenue de l'autre.

Si E_1 et E_2 sont indépendants :

$$P(E_1 \cap E_2) = P(E_2 \cap E_1) = P(E_1)\, P(E_2)$$

Si E_1, E_2 et E_3 sont indépendants deux à deux :

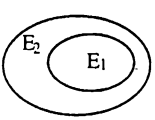
$$P(E_1 \cap E_2 \cap E_3) = P(E_1)\, P(E_2)\, P(E_3)$$

Si les événements E_1 et E_2 sont indépendants ;

alors $\bar{E}_1$ et E_2 sont indépendants
E_1 et $\bar{E}_2$ sont indépendants
$\bar{E}_1$ et $\bar{E}_2$ sont indépendants

■ ***Remarque 1 :***

Si $E_1 \subset E_2$



alors $P(E_1 \cap E_2) = P(E_1)$

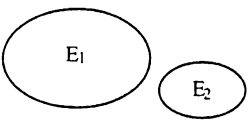
et $P(E_2 / E_1) = \frac{P(E_1 \cap E_2)}{P(E_1)} = 1$

et $$P(E_1 / E_2) = \frac{P(E_1 \cap E_2)}{P(E_2)} = \frac{P(E_1)}{P(E_2)}$$

Ainsi E_1 et E_2 ne sont pas indépendants.

Remarque 2 :

Si $E_1 \cap E_2 = \varnothing$



alors $P(E_1 \cap E_2) = P(\varnothing) = 0$

et $$P(E_1 / E_2) = \frac{P(E_1 \cap E_2)}{P(E_2)} = \frac{0}{P(E_2)} = 0$$

E_1 et E_2 ne sont pas indépendants.

1.5. Arbre des probabilités

Selon la formule des probabilités conditionnelles, on a :

$$P(A \cap B) = P(A / B)\, P(B)$$

que l'on peut généraliser pour les événements quelconques A_1, ..., A_n, selon le théorème de la multiplication, par :

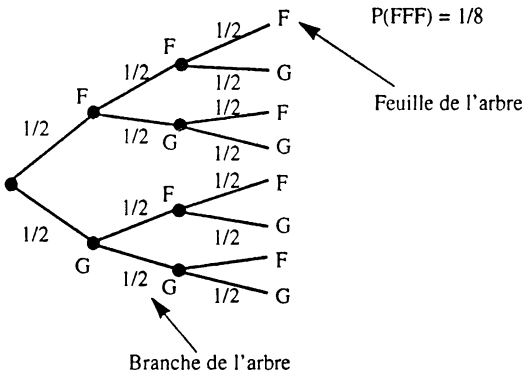
$P(A_1 \cap A_2 \cap A_3 \ldots \cap A_n) = P(A_1)\ P(A_2 / A_1)\, P(A_3 / A_1 \cap A_2) \ldots P(An / A_1 \cap A_2 \ldots \cap A_{n-1})$

Lors du calcul de probabilités sur une série d'expériences, il est parfois utile de représenter les cas possibles par un arbre afin de déduire plus facilement les résultats. La probabilité qu'un chemin particulier de l'arbre se produise (soit la probabilité de chaque feuille de l'arbre) correspond d'après le théorème de la multiplication au produit des probabilités de chaque branche du chemin.

• Exemple 1 :

En théorie, la probabilité d'avoir un garçon est de 1/2, de même que la probabilité d'avoir une fille. Quelle est la probabilité pour un couple ayant trois enfants d'avoir eu trois filles ?

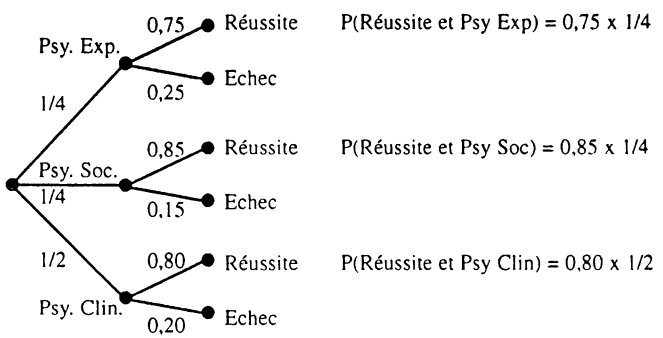
Ce calcul peut être déduit de l'arbre suivant :



Il y a huit possibilités d'obtenir une famille de trois enfants et un seul cas favorable pour qu'une famille ait trois filles sur l'ensemble des possibilités. Ainsi, la probabilité pour une famille de trois enfants d'avoir trois filles est de 1/8.

• *Exemple 2 :*

Dans une université, en maîtrise de psychologie le taux de réussite des étudiants ayant choisi une des trois sections, psychologie expérimentale, psychologie sociale ou psychologie clinique est respectivement de 0,75 ; 0,85 et 0,80 (ces données sont fictives). On sait que 1/4 des étudiants ont choisi la discipline psychologie expérimentale, 1/4 ont choisi la psychologie sociale et 1/2 ont choisi la psychologie clinique. Quelle est la probabilité pour qu'un étudiant pris au hasard obtienne sa maîtrise de psychologie dans cette université ?
Ce calcul peut être déduit de l'arbre suivant :



La probabilité de réussite en maîtrise dans cette université est égale à la somme des probabilités d'être reçu pour tout chemin de l'arbre aboutissant à une réussite, soit :
$P(R) = P(R \cap \text{Psy. Exp.}) + P(R \cap \text{Psy. Soc.}) + P(R \cap \text{Psy. Clin.})$
$P(R) = 0{,}75 \times 1/4 + 0{,}85 \times 1/4 + 0{,}80 \times 1/2 = 0{,}80$

1.6. Théorème de Bayes

Soit A et B deux événements, selon les **probabilités conditionnelles** on a :

$$P(A / B) = \frac{P(A \cap B)}{P(B)} \quad \text{soit} \quad P(A \cap B) = P(A / B)\, P(B)$$

et $$P(B / A) = \frac{P(A \cap B)}{P(A)} \quad \text{soit} \quad P(A \cap B) = P(B / A)\, P(A)$$

alors $P(A / B)\, P(B) = P(B / A)\, P(A)$ soit $P(A / B) = \dfrac{P(B / A)\, P(A)}{P(B)}$

or $P(B) = P(A \cap B) + P(\overline{A} \cap B)$

soit $P(B) = P(B / A)\, P(A) + P(B / \overline{A})\, P(\overline{A})$

Ainsi selon le théorème de Bayes :

$$P(A / B) = \frac{P(B / A)\, P(A)}{P(B / A)\, P(A) + P(B / \overline{A})\, P(\overline{A})}$$

On en déduit le modèle probabiliste de la démarche diagnostique :

On dit qu'un signe S est diagnostique de la maladie M si sa probabilité chez les malades, P(S / M), est différente de sa probabilité chez les non malades, $P(S / \overline{M})$. Ces deux probabilités (ou fréquences) ainsi que la probabilité de la maladie, P(M), sont connues d'après les statistiques médicales.

La probabilité qu'un sujet a d'être malade s'il porte le signe S, P(M/S), est déduite grâce au théorème de Bayes :

$$P(M / S) = \frac{P(S / M)\, P(M)}{P(S)} \quad \text{soit} \quad P(M / S) = \frac{P(S / M)\, P(M)}{P(S / M)\, P(M) + P(S / \overline{M})\, P(\overline{M})}$$

■ *Définitions*

Soit les événements M : être malade, $\overline{M}$: ne pas être malade, S : signe diagnostic positif (présence du signe S), et $\overline{S}$: signe diagnostic négatif (absence du signe S).

P(M) correspond à la *prévalence* instantanée de la maladie M. La prévalence instantanée d'une maladie ou de tout autre événement correspond au nombre de cas enregistrés dans une population déterminée, au moment de l'étude.

P(S/M) correspond à la *sensibilité* du signe S pour la maladie M (notée Se), c'est-à-dire à la capacité du signe diagnostique de détecter la maladie lorsqu'elle est présente. C'est le taux de vrais positifs chez les malades.

P($\overline{S}$/$\overline{M}$) correspond à la *spécificité* du signe S pour la maladie M (notée Sp), c'est-à-dire à la capacité de l'absence du signe diagnostique de déduire l'absence de maladie lorsqu'elle est réellement absente. C'est le taux de vrais négatifs chez les non malades.

P(M/S) correspond à la *valeur prédictive positive* du signe diagnostique S (notée VPP), c'est-à-dire à la proportion de sujets malades, chez ceux présentant le signe diagnostique S.

P($\overline{M}$/$\overline{S}$) correspond à la *valeur prédictive négative* du signe diagnostique S (notée VPN), c'est-à-dire à la proportion de sujets non malades, chez ceux ne présentant pas le signe diagnostique S.

P(S/$\overline{M}$) correspond à la *proportion de faux positifs*, c'est-à-dire à la proportion de sujets présentant le signe diagnostique S, chez ceux qui ne sont pas malades.

P($\overline{S}$/M) correspond à la *proportion de faux négatifs*.

P($\overline{S}$/$\overline{M}$) correspond à la *proportion de vrais négatifs*, c'est-à-dire à la poportion de sujets ne présentant pas le signe diagnostique S, chez ceux qui ne sont pas malades.

Ces quatre dernières définitions peuvent être résumées dans le tableau suivant :

Probabilité →	+ Présence du signe diagnostique	− Absence du signe diagnostique
+ Malade	P(M / S) VPP	P(M / $\overline{S}$)
− Non malade	P($\overline{M}$ / S)	P($\overline{M}$ / $\overline{S}$) VPN

Tableau I

• Exemple 1 :

Dans une population la prévalence d'une pathologie P est de 40 %. La spécificité du signe diagnostique S, pour cette pathologie P, est de 0,5 et sa sensibilité est de 0,7.
Quelle est la valeur prédictive positive du signe diagnostique S ?

Soit P : présenter la pathologie P ; $\overline{P}$: ne pas présenter la pathologie P ; S : signe diagnostic positif (présence du signe S) ; et $\overline{S}$: signe diagnostic négatif (absence du signe S).
P(P) = 40 % = 0,4
Sp = P($\overline{S}$ / $\overline{P}$) = 0,5
Se = P(S/P) = 0,7

$$VPP = P(P / S) = \frac{P(S / P)\,P(P)}{P(S / P)\,P(P) + P(S / \overline{P})\,P(\overline{P})}$$

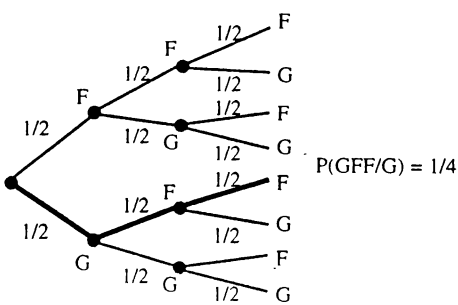
$$VPP = \frac{0,7 \times 0,4}{0,7 \times 0,4 + 0,5 \times 0,6} = \frac{0,28}{0,58} = 0,4828 = 48,28\ \%$$

Cette probabilité correspond à la probabilité de présenter la pathologie P lorsque le signe diagnostique est positif.

Le calcul de probabilités conditionnelles revient à former un arbre inversé. En effet, sachant que nous nous trouvons sur une certaine branche de l'arbre, quelle est la probabilité de l'événement en question ?

• Exemple 2 :

Reprenons l'arbre précédent (paragraphe 1.5.). Quelle est la probabilité pour une famille de trois enfants d'avoir un garçon puis deux filles sachant qu'elle a d'abord eu un garçon ?



Il y a quatre cas possibles pour une famille de trois enfants ayant d'abord eu un garçon, et un seul cas favorable pour que cette famille ait ensuite deux filles. Ainsi, la probabilité pour une famille de trois enfants d'avoir deux filles après avoir eu un garçon est 1/4.

2. Introduction aux statistiques

2.1. Variable aléatoire, fonction de répartition

■ **Une *variable aléatoire* X, définie sur un ensemble fondamental F, est une fonction de F sur l'ensemble des nombres réels (IR), telle que l'image inverse de tout intervalle de IR soit un événement de F. L'application F → IR s'appelle variable aléatoire.**

Tout sous-ensemble de F étant un événement, toute fonction à valeur réelle définie sur F est une variable aléatoire (voir la distinction entre variable qualitative ou quantitative dans le chapitre I).

Soit X et Y des variables aléatoires définies sur le même ensemble fondamental F, les fonctions suivantes sont aussi des variables aléatoires :

$(X+Y)(f) = X(f) + Y(f)$
$(kX)(f) = kX(f)$ avec k une constante quelconque
$(X + k)(f) = X(f) + k$
$(XY)(f) = X(f)\,Y(f)$

▒ *Notation :*

Soit X la variable aléatoire associée à l'épreuve E et x un nombre réel fixé.

$P(X=x)$ désigne la probabilité de l'événement : X a pour valeur x.

$P(x_1 \leq X \leq x_2)$ désigne la probabilité de l'événement : X a une valeur comprise entre x_1 et x_2.

■ **La loi de X est définie par sa *fonction de répartition*.**

La fonction de répartition F(x) associée à la variable aléatoire X est définie par $F(x) = P(X < x)$.

F est continue à gauche, jamais décroissante.

$F(-\infty) = 0$ c'est-à-dire que la probabilité pour qu'un nombre quelconque soit inférieur à $-\infty$ est zéro ;

et $F(+\infty) = 1$ c'est-à-dire qu'un nombre est toujours inférieur à $+\infty$.

Soit X une variable aléatoire définie sur un ensemble dénombrable de valeurs. Soit $x_1, x_2, ..., x_n, ...$ ces valeurs.

La loi de X est définie par les probabilités p_i des nombres x_i, c'est-à-dire $p_i = P(X=x_i)$.

On a : $0 \leq p_i \leq 1$ et $\sum_{i=1}^{\infty} p_i = 1$

■ **L'espérance mathématique E(X), c'est-à-dire la moyenne de X, est définie par :**

$$E(X) = \mu = \sum_{i=1}^{\infty} p_i \, x_i$$

■ **La variance de X est définie par :**

$$Var(X) = \sigma^2 = \sum_{i=1}^{\infty} p_i \, (x_i - \mu)^2$$

soit $$\sigma^2 = \left(\sum_{i=1}^{\infty} p_i \, x_i^2 \right) - \mu^2 = E(X^2) - [E(X)]^2$$

μ est une caractéristique de position de la loi de X ; σ^2 est une caractéristique de dispersion des valeurs individuelles x_i que peut prendre X autour de μ.

▩ ***Opérations sur les variables aléatoires :***

Si $Y = aX + b$ $\quad E(Y) = a\,E(X) + b$
$\quad\quad\quad

2.2. Densité de probabilité d'une variable aléatoire continue

Soit X une variable aléatoire définie sur un ensemble continu de valeurs réelles. On dit que X est une variable aléatoire absolument **continue si la probabilité que X soit égale à une valeur particulière x est nulle :** $P(X=x) = 0, \forall x \in IR$; seule la probabilité de X de tomber dans un intervalle de valeurs est définie.

On appelle densité de probabilité (d.d.p.) la fonction f(x) définie par : $P(x \leq X < x + dx) = f(x)\,dx$ quand $dx \to 0$.

La fonction f(x) est une densité de probabilité si et seulement si :

1) $f(x) \geq 0$

2) $\int_{-\infty}^{+\infty} f(x)dx = 1$

3) $P(X \in [a,b]) = \int_{a}^{b} f(x)dx$

La moyenne ou espérance mathématique de X est définie par :

$$E(X) = \mu = \int_{-\infty}^{+\infty} x\, f(x)dx$$

La variance de X est définie par :

$$Var(X) = \sigma^2 = \int_{-\infty}^{+\infty} (x-\mu)^2\, f(x)dx$$

soit

$$\sigma^2 = \int_{-\infty}^{+\infty} x^2\, f(x)dx - \mu^2$$

2.3. Statistiques de position : Moyenne, Médiane, Mode

La moyenne, le mode et la médiane sont des indices de position, valeurs caractéristiques ou représentatives d'un ensemble de données. Ces indices sont soit des statistiques caractérisant un échantillon, soit des paramètres

caractérisant une population. **Les statistiques sont notées par les lettres de l'alphabet, et les paramètres par des lettres grecques.**

2.3.1. La moyenne

La moyenne (*mean* en anglais) ou moyenne arithmétique d'un caractère quantitatif (ensemble d'observations quantitatives) est le quotient de la somme de toutes les valeurs observées de ce caractère par le nombre total des observations. Elle est le plus souvent notée m ou $\bar{x}$. Elle est notée μ lors du calcul de la moyenne d'un caractère observé sur l'ensemble d'une population (théorique de référence).

La moyenne :

$$m = \bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \sum_{i=1}^{n} \frac{x_i}{n}$$

que l'on note aussi :

$$m = \sum \frac{x}{n} = \frac{\text{somme des observations}}{\text{nombre d'observations}}$$

• *Exemple :*

La moyenne arithmétique des nombres 4, 6, 9, 2, 15 est

$$m = \frac{4+6+9+2+15}{5} = \frac{36}{5} = 7{,}2$$

Dans le cas de données groupées où les mêmes données se retrouvent plusieurs fois, c'est-à-dire lorsque plusieurs observations correspondent à la même modalité du caractère étudié, tel que : 2, 2, 2, 2, 2, 5, 5, 5, 5, 6, 6, 6, 9, 9, 9, 9, 9, 9, 9 on utilise la formule suivante :

$$m = \bar{x} = \frac{n_1x_1 + n_2x_2 + \ldots + n_nx_n}{n_1 + n_2 + \ldots + n_n} = \frac{\sum_{i=1}^{n} n_ix_i}{\sum_{i=1}^{n} n_i} = \sum_{i=1}^{n} \frac{n_ix_i}{N}$$

avec $N = \sum_i n_i$

ou encore

$$m = \frac{\sum(\text{nombre d'observations de la modalité} \times \text{valeur de la modalité})}{\text{nombre total d'observations}}$$

soit $$m = \frac{(5\times 2)+(4\times 5)+(3\times 6)+(7\times 9)}{5+4+3+7} = \frac{111}{19} = 5{,}84$$

Dans le cas de *variables regroupées par classes*, si les classes sont de même amplitude, on attribue à chaque observation la valeur centrale de sa classe et on calcule la moyenne grâce à la formule suivante :

$$m = \frac{\sum(\text{nombre d'observations de la classe} \times \text{valeur centrale de la classe})}{\text{nombre total d'observations}}$$

2.3.2. La médiane

La médiane (*median* en anglais) d'un ensemble de nombres ordonnés de façon croissante ou décroissante est la **valeur centrale** (dans le cas d'un nombre impair d'observations), ou **la moyenne arithmétique des deux valeurs centrales** (dans le cas d'un nombre pair d'observations). Elle est souvent notée M.

La médiane correspond à la modalité qui partage l'ensemble des observations classées (dans l'ordre croissant ou décroissant) en **deux moitiés de tailles égales**, contenant donc chacune 50 % des observations (50 % des observations avant la médiane et 50 % après). Ainsi, dans le cas de lois symétriques autour de la moyenne (telles que la loi normale), la médiane est égale à la moyenne.

Dans le cas d'une variable discontinue, la médiane n'existe pas toujours. Dans le cas continu, si la valeur médiane appartient à une classe, elle peut être déterminée graphiquement, ou par interpolation linéaire qui présuppose l'équirépartition à l'intérieur d'une classe.

• *Exemple 1 :*

Dans le cas d'un *nombre impair d'observations*, la médiane des nombres suivants 2, 2, 4, 7, 9, 10, 12, 13, 13 est 9.

La médiane est alors la modalité de l'observation de rang :
$\frac{n+1}{2}$ avec n correspondant au nombre total d'observations.

• Exemple 2 :

Dans le cas d'un *nombre pair d'observations*, la médiane des nombres suivants 2, 2, 4, 7, 9, 9, 10, 12, 13, 13 est :

$$M = \frac{9+9}{2} = 9$$

La médiane est alors la moyenne arithmétique des deux valeurs centrales, c'est-à-dire des deux observations respectivement de rang :
$\frac{n}{2}$ et $\frac{n}{2}+1$ avec n correspondant au nombre total d'observations.

2.3.3. Le mode

Le mode ou la dominante (*dominant* en anglais) est la valeur ou la classe qui apparaît le plus souvent dans la distribution de la variable étudiée, c'est la **modalité qui a la plus grande fréquence**. Le ou les mode(s) peuvent être déterminés graphiquement, ils correspondent aux valeurs les plus observées, c'est-à-dire graphiquement aux points, ou aux classes, dont les bâtons de diagrammes ou des histogrammes sont maximum et forment des pics.

Le mode ou la dominante d'un caractère **discontinu** est la modalité qui a le plus fort **effectif**. Le mode d'un caractère **continu** est la modalité qui a la plus forte **densité**. C'est la valeur la plus probable d'un caractère. Dans le cas de caractères continus, le mode est sensible au choix du regroupement par classe.

De plus, le regroupement des classes peut faire apparaître des modes là où il n'y en a pas. Il faut alors être attentif à la formation de la surface du bâton d'un histogramme qui doit être proportionnelle à l'effectif des observations dans cette classe, et la base des bâtons (des classes) doit être proportionnelle à l'étendue de la classe.

Dans le cas où plusieurs modes apparaissent on parle alors de **distribution multimodale**. Par exemple, dans le cas de deux modes on parle de distribution bimodale. Dans le cas où l'on observe le même effectif pour chaque modalité, et qu'il n'y a donc pas de mode apparent, alors **la distribution est dite amodale**.

Dans le cas de courbes unimodales symétriques, telles que la loi normale (que nous abordons plus tard), la moyenne, le mode et la médiane coïncident.

• Exemple 1 :

Le mode de l'ensemble des observations suivantes 2, 2, 4, 7, 9, 10, 12, 13, 13, 13 est 13.

• *Exemple 2 :*

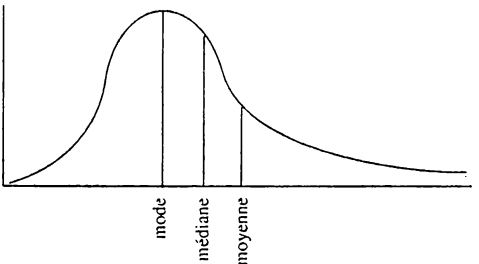
Le mode de l'ensemble des observations suivantes 2, 2, 4, 7, 9, 10, 10, 10, 10, 12, 13, 13 est 10.

• *Exemple 3 :*

Les modes de l'ensemble des observations suivantes 2, 2, 2, 2, 4, 7, 9, 10, 10, 10, 10, 12, 13, 13 sont 2 et 10.

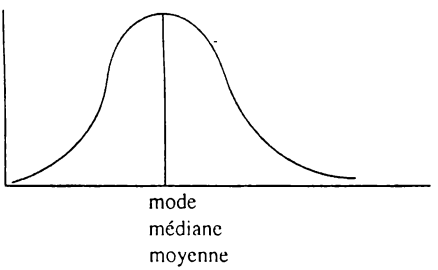
• *Exemple 4 :*

Positions de la moyenne, de la médiane et du mode dans le cas d'une courbe asymétrique unimodale :



• *Exemple 5 :*

Les positions de la moyenne, de la médiane et du mode sont équivalentes dans le cas d'une courbe symétrique unimodale :



Sur un échantillon observé, on voit que les statistiques (mode, médiane et moyenne) caractérisent une distribution **symétrique** quand elles sont **égales, asymétrique à gauche** quand le mode est **inférieur à la moyenne**, et à droite quand le mode est **supérieur à la moyenne**. Ainsi, les statistiques donnent des informations sur l'échantillon observé.

2.4. Statistiques de dispersion : Etendue, Ecart-moyen, Ecart-type, Variance, Covariance

En sciences humaines la variabilité des observations est courante, elle traduit la fluctuation d'un caractère d'un individu à l'autre, ou d'un individu d'un moment à l'autre et pas seulement la fluctuation liée à l'erreur de mesure. Ainsi, la distribution d'une variable ne peut pas être uniquement représentée par son indice de position (sa moyenne). En effet, des distributions différentes peuvent avoir un même indice de position. Les indices de dispersion (notamment l'écart-moyen) permettent alors de représenter les distances entre les différentes observations. Finalement, ce qui représente le mieux une distribution c'est sa moyenne, l'intervalle de confiance autour de cette moyenne, et l'étendue du caractère observé.

2.4.1. L'étendue

L'étendue (*range* en anglais), ou *intervalle de variation,* d'un ensemble de mesures observées est la différence entre le plus grand et le plus petit nombre observé. Le plus souvent, on note l'étendue par le couple formé par le plus petit nombre puis le plus grand nombre observé du caractère étudié.

• Exemple 1 :

L'étendue de l'ensemble des observations suivantes : 4, 8, 9, 9, 15, 18, 25
est 25 - 4 = 21
que l'on note : 4 - 25

• Exemple 2 :

L'étendue de l'ensemble des observations suivantes : 0, 2, 2, 5, 6, 8, 8
est 8 - 0 = 8
que l'on note : 0 - 8

2.4.2. L'écart-moyen

L'écart-moyen, ou *écart à la moyenne,* est la moyenne de l'ensemble des distances entre chaque observation et la moyenne arithmétique des observations. Cet indice de dispersion permet d'exprimer les divers écarts entre les observations et la moyenne en terme de distance, c'est-à-dire à partir des valeurs absolues des différences entre les observations et la moyenne, des $|x_i - \bar{X}|$. L'écart-moyen du caractère X est défini par :

$$\text{Ecart moyen} = \text{E.M.} = \frac{\sum_{i=1}^{n} |x_i - \bar{x}|}{n}$$

avec $\bar{x}$ la moyenne arithmétique des observations, $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$.

• *Exemple :*

L'écart-moyen de l'ensemble des observations suivantes 4, 6, 6, 7, 8, 9, 9, 11, 15, 16, 18, 20

est $$\text{E.M.} = \frac{\sum_{i=1}^{n} |x_i - \bar{x}|}{n} = \frac{\sum_{i=1}^{12} |x_i - 10,75|}{12} = 4,333$$

car

$$\text{E.M.} = \frac{|4-10,75| + 2|6-10,75| + |7-10,75| + |8-10,75| + 2|9-10,75| + |11-10,75| + |15-10,75| + |16-10,75| + |18-10,75| + |20-10,75|}{12}$$

avec $$\bar{x} = \frac{\sum_{i=1}^{12} x_i}{12} = \frac{4+6+6+7+8+9+9+11+15+16+18+20}{12} = 10,75$$

2.4.3. L'écart-type

L'écart-type (*standard deviation* en anglais), ou *écart quadratique moyen*, est la moyenne quadratique de l'ensemble des écarts entre chaque observation et la moyenne arithmétique des observations. **C'est la racine carrée de la variance**. Avec la variance, l'écart-type est l'indice de dispersion le plus utilisé, notamment lors de l'utilisation du modèle de la loi normale.

L'écart-type du caractère X étudié, calculé sur l'ensemble des données d'une population, est défini par :

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n}}$$

• *Exemple :*

L'écart-type de l'ensemble des données suivantes :
6, 6, 7, 9, 9, 11, 16, 16, 20, 20

est $$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}} = \sqrt{\frac{\sum_{i=1}^{10}(x_i - 12)^2}{10}}$$

soit $$\sigma = \sqrt{\frac{2(6-12)^2 + (7-12)^2 + 2(9-12)^2 + (11-12)^2 + 2(16-12)^2 + 2(20-12)^2}{10}} = 5.25$$

L'écart-type estimé d'un caractère X étudié d'une population, estimé à partir des observations d'un échantillon de taille n issu de cette population de référence, est noté s et est défini par :

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

• *Exemple :*

Si l'on considère les données suivantes comme celles d'un échantillon représentatif de la population de référence : 6, 6, 7, 9, 9, 11, 16, 16, 20, 20
alors l'écart-type, du caractère X étudié de la population de référence, estimé à partir des données de l'échantillon, est :

$$s = \sqrt{\frac{\sum_{i=1}^{10}(x_i - 12)^2}{10-1}} = \sqrt{\frac{276}{9}} = 5,54$$

2.4.4. La variance

La variance (*variance* en anglais), ou *carré moyen*, est le carré de l'écart-type. Pour un caractère X étudié sur l'ensemble d'une population, la variance est définie par :

$$\sigma^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}$$

ou encore $$\sigma^2 = \frac{\sum_{i=1}^{n} x_i^2}{n} - \bar{x}^2$$

soit
$$\sigma^2 = \frac{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}}{n}$$

Puis, dans le cas où l'on retrouve plusieurs fois les mêmes valeurs x_i :

$$\sigma^2 = \frac{\sum_{i=1}^{n} n_i (x_i - \mu)^2}{\sum_{i=1}^{n} n_i}$$

• *Exemple :*

La variance de l'ensemble des données suivantes 6, 6, 7, 9, 9, 11, 16, 16, 20, 20

est
$$\sigma^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n} = \frac{\sum_{i=1}^{10} (x_i - 12)^2}{10}$$

soit
$$\sigma^2 = \frac{2(6-12)^2 + (7-12)^2 + 2(9-12)^2 + (11-12)^2 + 2(16-12)^2 + 2(20-12)^2}{10} = 27.6$$

ou encore
$$\sigma^2 = \frac{\sum_{i=1}^{n} x_i^2}{n} - \bar{x}^2 = \frac{\sum_{i=1}^{10} x_i^2}{10} - 12^2$$

soit
$$\sigma^2 = \frac{2\times 6^2 + 7^2 + 2\times 9^2 + 11^2 + 2\times 16^2 + 2\times 20^2}{10} - 12^2 = \frac{1716}{10} - 144 = 27.6$$

La variance estimée d'un caractère X étudié d'une population, estimée à partir des observations d'un échantillon de taille n issu de cette population, est noté s^2 et est défini par :

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$$

Si l'on considère les données suivantes comme celles d'un échantillon représentatif de la population de référence : 6, 6, 7, 9, 9, 11, 16, 16, 20, 20

alors la variance estimée du caractère X étudié est :

$$s^2 = \frac{\sum_{i=1}^{10} (x_i - 12)^2}{10 - 1} = \frac{276}{9} = 30{,}67$$

2.4.5. Le coefficient de variation

Le coefficient de variation (CV) permet de mesurer la représentativité de la moyenne arithmétique x par rapport à l'ensemble des données. Il est défini par le pourcentage suivant :

$$CV = \frac{s}{\bar{x}} \times 100$$

Le coefficient de variation permet d'**apprécier l'homogénéité de la distribution.** Si CV < 15 % alors on conclut à une bonne homogénéité de la distribution.

Le coefficient de variation ne dépend pas des unités choisies. Il permet de comparer deux distributions, dans le cas où les deux séries de données ne sont pas exprimées avec la même unité, ou dans le cas où les deux séries de données sont de moyennes arithmétiques très différentes.

2.4.6. La covariance

La covariance (*covariance* en anglais) mesure la variation simultanée de deux variables X et Y de même effectif. Elle mesure le **degré de dépendance de X et Y**, et elle est définie par :

$$\text{Covar}(X,Y) = \sigma_{x,y} = E(XY) - \mu_x \mu_y$$

soit $\sigma_{x,y} = E(X - \mu_x)(Y - \mu_y)$

$$\sigma_{X.Y} = \frac{\sum_{i=1}^{n} (X_i - \mu_X)(Y_i - \mu_Y)}{n}$$

qui peut être estimée par :

$$s_{X,Y} = \frac{\sum_{i=1}^{n} (X_i - \overline{X})(Y_i - \overline{Y})}{n-1}$$

Si les variables X et Y sont indépendantes, alors leur degré de dépendance est nul, c'est-à-dire : $\sigma_{x,y} = 0$

2.4.7. La corrélation

Une extension de la covariance, qui ne tient pas compte de l'unité de mesure de X et Y, est la ***corrélation*** entre X et Y qui est toujours comprise entre -1 et +1, et est définie par :

$$\rho = \frac{\sigma_{X,Y}}{\sigma_X \sigma_Y}$$

La corrélation est définie plus en détail dans le chapitre X.

3. L'échantillonnage d'une population

Une population correspond à un ensemble d'individus ou d'objets de même nature. Lorsque la taille d'une population est très grande, voire infinie, il est le plus souvent impossible d'en connaître exhaustivement les caractéristiques. C'est pourquoi on étudie le plus souvent une population grâce à la constitution d'un sous-ensemble, c'est-à-dire d'un échantillon issu de cette population, dite de référence.

Pour que l'on puisse ensuite extrapoler à l'ensemble de la population les caractéristiques de l'échantillon, il faut que celui-ci soit *représentatif*. Si l'échantillon est constitué de presque tous les individus de la population, on aura confiance dans sa représentativité : c'est ce qui arrive à l'occasion des recensements de la population française ; la moyenne d'âge obtenue est très proche de celle de la population totale.

On conçoit intuitivement que si l'on utilise un échantillon de seulement 10 sujets, on a des chances que la moyenne d'âge obtenue s'écarte de la moyenne vraie (théorique) de la population française. Plus on augmente la taille de l'échantillon, plus on se rapproche de la population française, mais plus cela est coûteux, il faut donc déterminer *a priori* la taille de l'échantillon. Cependant, même quand cet échantillon est suffisamment grand, il peut y avoir un biais de recrutement. S'il est par exemple constitué dans les écoles, il ne sera pas représentatif pour le caractère âge.

3.1. Randomisation

Pour estimer les caractéristiques d'une population, il faut donc un échantillon de taille suffisante et sans biais de représentativité. La formation d'un échantillon représentatif d'une population de référence consiste en ***un tirage au sort*** (tirage aléatoire ou randomisation) d'un sous-ensemble de sujets parmi l'ensemble de tous les sujets de la population. La plupart des tirages au sort présupposent que **chaque sujet a la même probabilité d'être sélectionné**. Une méthode de tirage au sort possible consiste à attribuer un numéro à chacun des sujets de la population considérée, de taille N (numéros de 1 à N), puis en actionnant n fois (n correspondant à la taille n souhaitée de l'échantillon à constituer, $n < N$) la fonction de randomisation (du logiciel Excel par exemple ou d'un langage informatique), les numéros notés sélectionnés aléatoirement par l'ordinateur (ou à partir d'une table de nombre au hasard) permettront de former théoriquement un échantillon représentatif de la population initiale.

La constitution des échantillons joue un rôle important dans l'interprétation des résultats, puisqu'elle est à la base de l'extrapolation de ces résultats. ***C'est l'échantillon que nous observons, mais c'est la population que nous voulons connaître.*** Un échantillon constitué sans tirage au sort peut plus facilement introduire des biais liés à des facteurs de confusion (ou facteurs exogènes non contrôlés) qui fausseront les interprétations. Prenons l'exemple de l'étude de la liaison entre les deux caractères "survenue d'un cancer du poumon" et "catégorie socio-économique". La conclusion affirmant que certaines catégories C_1 et C_2 présentent davantage de risques de survenue d'un cancer sera biaisée si le facteur de confusion "consommation de tabac" n'est pas contrôlé. En effet, si du fait de la constitution des échantillons, les catégories C_1 et C_2 comprennent davantage de consommateurs de tabac (ce qui augmenterait le risque), au lieu d'attribuer la recrudescence de cancers à la consommation de tabac, on aurait tendance à penser qu'elle est liée à la catégorie socio-économique.

C'est pourquoi, quand cela est possible, la constitution des échantillons par tirage au sort permet de maintenir les facteurs de confusion constants, évite les biais, et permet une interprétation plus fine des résultats. Dans les cas où la constitution des échantillons par tirage au sort n'est pas possible, on a recours à des méthodes plus sophistiquées, telles que, par exemple, l'analyse de régression multiple qui permet, mais de façon non systématique, d'identifier les facteurs de confusion.

3.2. Echantillons appariés

Dans le but d'obtenir des interprétations fiables, un autre moyen de constituer des échantillons comparables, en maintenant constantes la plupart des variables exogènes, est d'apparier les échantillons. Il s'agit notamment de comparer les sujets à eux-mêmes dans des situations différentes (par exemple à deux temps différents ou selon deux traitements différents). Les études des comportements des jumeaux permettent aussi de contrôler la plupart des variables exogènes. Par ailleurs, en psychologie expérimentale, on peut être amené à constituer des échantillons de souris appariés sur une ou plusieurs variables. Par exemple, on peut apparier deux échantillons de souris, l'un traité par A et l'autre traité par B, sur les deux variables "taille" et "poids" ; c'est-à-dire que les souris de même taille et même poids seront comparées les unes aux autres, alors qu'elles auront été traitées différemment par A ou par B.

3.3. Cas-témoin

Lorsque l'on compare par exemple l'effet d'un traitement chez deux groupes de sujets, dont l'un est traité et l'autre non, on parle alors d'études cas-témoin (les témoins dans ce cas sont malades comme les cas, seul le traitement diffère), avec introduction d'un groupe témoin ou contrôle. Pour vérifier l'efficacité du traitement, il est non seulement souhaitable que ces deux groupes soient constitués par randomisation, mais aussi que le traitement soit administré en aveugle. Les sujets sont "aveugles" en ce qui concerne le traitement, ils ne savent pas s'ils reçoivent le traitement ou un placebo (groupe non traité ou traité par une substance neutre telle que de l'eau distillée). Enfin, pour s'assurer qu'aucun facteur de confusion n'intervient, il est également souhaitable que l'administration du traitement ou du placebo n'influence pas l'expérimentateur dans ses observations, et qu'il ne distingue pas les sujets traités de ceux sous placebo : on parle alors d'études en double-aveugle. Ce type d'étude peut être réalisée quand on compare deux traitements apparemment identiques dans leur forme galénique, ou dans leurs procédures thérapeutiques.

4. Lois de probabilité

4.1. Loi binomiale

Soient les épreuves répétées et indépendantes d'une même expérience n'ayant que deux résultats possibles (succès et échec). Par exemple, dans le cas d'une urne de dimension infinie contenant une proportion p de boules

blanches et q =1-p de boules rouges, si on suppose que l'événement "tirer au sort une boule blanche" est appelé succès et l'événement "tirer une boule rouge" est appelé échec, alors p est la probabilité de l'événement succès et q est la probabilité de l'événement échec.

Si on tire au sort n boules de l'urne avec remise (épreuves répétées et indépendantes), **la loi de probabilité du nombre X de succès** (c'est-à-dire de boules blanches) **est la loi binomiale :**

$$P(X = k) = C_n^k \, p^k \, q^{n-k}$$

avec $0 \leq X \leq n$ et $q = 1-p$

ou encore $$P(X = k) = \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k}$$

La probabilité de n'obtenir aucun succès, c'est-à-dire une boule rouge pour chacun des n tirages au sort est q^n. La probabilité d'obtenir n succès, c'est-à-dire une boule blanche pour chacun des n tirages au sort est p^n.

La distribution binomiale est de moyenne théorique $\mu = np$,
de variance $\sigma^2 = npq$
et d'écart-type $\sigma = \sqrt{npq}$.

• Exemple :

On tire au sort 10 personnes dans une population de 100 personnes composée d'autant d'hommes que de femmes.

1. Quelle est la probabilité que cet échantillon de 10 personnes soit composé exactement de 4 femmes ?

$P(X = 4) = C_{10}^4 \, p^4 \, q^6$

soit $$P(X = 4) = \frac{10 \times 9 \times 8 \times 7}{4 \times 3 \times 2} \left(\frac{1}{2}\right)^4 \left(\frac{1}{2}\right)^6 = 210 \times \frac{1}{16} \times \frac{1}{64} = 0{,}205$$

2. Quelle est la probabilité que cet échantillon de 10 personnes soit composé d'au moins une femme ?

$$P(X = 0) = C_{10}^0 \, p^0 \, q^{10} = \left(\frac{1}{2}\right)^{10} = \frac{1}{2^{10}} = \frac{1}{1024} = 9{,}77 \times 10^{-4}$$

P(au moins une femme) = 1-P(X = 0) = 1 - $9{,}77 \times 10^{-4}$ = 0,999

4.2. Loi de Poisson

La distribution binomiale correspond à une distribution de Poisson quand une des éventualités a une probabilité très faible de survenir, quand p ou q > 0,03 (voir 4.1.).

La variable aléatoire X, définie sur l'ensemble des nombres entiers (0, 1, 2,..., n,...), est distribuée selon la loi de Poisson de paramètre λ (avec $\lambda > 0$) si :

$$P(X = k) = \frac{e^{-\lambda} \lambda^k}{k!}$$

avec e = 2,71828... et λ une constante donnée.

La distribution de Poisson est de moyenne théorique $\mu = \lambda$,
de variance $\sigma^2 = \lambda$
et d'écart-type $\sigma = \sqrt{\lambda}$.

Remarque :

On peut montrer que la loi de Poisson tend vers la loi normale centrée réduite $\frac{X - \lambda}{\sqrt{\lambda}}$ quand λ tend vers l'infini.

• Exemple :

Si la probabilité pour que k sujets sur 1 000 soient atteints d'une pathologie P suit une distribution de Poisson $P(X = k) = \frac{e^{-1}(1)^k}{k!}$, alors la probabilité pour que 5 sujets sur 1 000 soient atteints de cette pathologie P est :

$$P(X = 5) = \frac{e^{-1}(1)^5}{5!} \quad \text{soit} \quad P(X = 5) = \frac{1}{e \times 120} = 3{,}07.10^{-3}$$

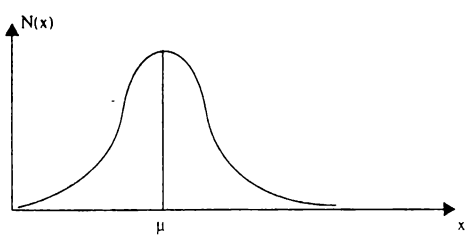
4.3. Loi normale

4.3.1. Formulation

La distribution normale de Laplace (1749-1827) et Gauss (1777-1855), dite loi normale ou loi gaussienne, est une distribution continue de densité :

$$N(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma^2}(x - \mu)^2\right)$$

La loi normale N(x), notée $N(\mu, \sigma^2)$, est une **fonction unimodale symétrique en forme de cloche** dont la représentation graphique est :



On dit que la variable aléatoire X est distribuée normalement ou selon une loi normale de moyenne μ et de variance σ^2. On note $X \rightarrow N(\mu,\sigma^2)$.

On a observé dans la nature que la **variabilité d'un grand nombre de mesures biologiques pouvait être modélisée par la loi normale**. Ainsi les histogrammes que l'on peut construire en mettant en abscisse les classes de valeurs et en ordonnée les fréquences avec lesquelles on tombe dans ces classes sont souvent proches de la courbe en cloche de Laplace-Gauss.

Un théorème mathématique démontre que l'addition d'une infinité de variables aléatoires de n'importe quelle loi, mais indépendantes entre elles, conduit à une variable aléatoire distribuée toujours normalement.

■ *Théorème central limite :*

La variable aléatoire **somme de n variables aléatoires indépendantes** suit une loi de probabilité qui converge vers une loi normale de moyenne $n\mu$ et de variance $n\sigma^2$. En pratique cette convergence est assurée **pour $n \geq 30$.**

■ *Pour les plus avertis, montrons que N(x) est une densité de probabilité (ddp) de moyenne m et de variance σ^2.*

Les calculs suivants permettent de démontrer que **la loi normale est une densité de probabilité**, mais ne sont pas nécessaires pour la compréhension de la suite de ce livre. N(x) est une ddp puisque N(x) est toujours positive (σ et la fonction exponentielle étant toujours positives) et la somme des probabilités vaut 1.

Soit $\int_{-\infty}^{+\infty} N(x)dx = 1$

$$\int_{-\infty}^{+\infty} N(x)dx = \int_{-\infty}^{+\infty} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right)dx = \int_{-\infty}^{+\infty} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\left(\frac{x-\mu}{\sigma\sqrt{2}}\right)^2\right)dx$$

on effectue un changement de variable et on pose :

$$t = \frac{x-\mu}{\sigma\sqrt{2}}$$

soit $x-\mu = \sigma\sqrt{2}\, t$

d'où $(x-\mu)' = (\sigma\sqrt{2}\, t)'$; soit $dx = \sigma\sqrt{2}\, dt$

Alors : $\int_{-\infty}^{+\infty} N(x)dx = \int_{-\infty}^{+\infty} \frac{1}{\sigma\sqrt{2\pi}} \exp(-t^2)\sqrt{2}\,\sigma\, dt$

soit $\int_{-\infty}^{+\infty} N(x)dx = \int_{-\infty}^{+\infty} \frac{1}{\sqrt{\pi}} \exp(-t^2)\, dt = \frac{1}{\sqrt{\pi}} \int_{-\infty}^{+\infty} \exp(-t^2)\, dt = 1$

sachant que $\int_{-\infty}^{+\infty} \exp(-t^2)\, dt = \sqrt{\pi}$.

On peut calculer *l'espérance et la variance* de la loi normale :

$$\boxed{E(x) = \int_{-\infty}^{+\infty} x\, N(x)\, dx = \mu}$$

car $E(x) = \int_{-\infty}^{+\infty} x\, N(x)\, dx = \int_{-\infty}^{+\infty} x \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\left(\frac{x-\mu}{\sigma\sqrt{2}}\right)^2\right) dx$

on pose :

$$t = \frac{x-\mu}{\sigma\sqrt{2}} \quad \text{soit} \quad x-\mu = \sigma\sqrt{2}\, t \quad \text{et} \quad x = \mu + \sigma\sqrt{2}\, t$$

d'où $(x-\mu)' = (\sigma\sqrt{2}\, t)'$ soit $dx = \sigma\sqrt{2}\, dt$

Alors :

$$E(x) = \int_{-\infty}^{+\infty} (\mu + \sigma\sqrt{2}\, t) \frac{1}{\sigma\sqrt{2\pi}} \exp(-t^2)\, \sqrt{2}\, \sigma\, dt$$

$$E(x) = \int_{-\infty}^{+\infty} (\mu + \sigma\sqrt{2}\, t) \frac{1}{\sqrt{\pi}} \exp(-t^2)\, dt$$

$$E(x) = \frac{1}{\sqrt{\pi}} \left(\int_{-\infty}^{+\infty} \mu \exp(-t^2)\, dt + \int_{-\infty}^{+\infty} \sigma\sqrt{2}\, t \exp(-t^2)\, dt \right)$$

$$E(x) = \frac{1}{\sqrt{\pi}} \left(\mu\sqrt{\pi} + \sigma\sqrt{2} \left[-\frac{1}{2} \exp(-t^2) \right]_{-\infty}^{+\infty} \right)$$

or $\left[-\frac{1}{2} \exp(-t^2) \right]_{-\infty}^{+\infty} \to 0$

donc $E(x) = \mu$

De même on montre que

$$\boxed{Var(x) = \int_{-\infty}^{+\infty} x^2\, N(x)\, dx = \sigma^2}$$

car dans le cas où $\mu=0$, c'est-à-dire dans le cas d'une loi normale centrée :

$$\text{Var}(x) = \int_{-\infty}^{+\infty} x^2 \, N(x)\,dx = \int_{-\infty}^{+\infty} x^2 \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{x^2}{2\sigma^2}\right) dx$$

on pose :

$$t = \frac{x}{\sigma\sqrt{2}} \qquad \text{soit} \quad x = \sigma\sqrt{2}\, t$$

d'où $(x)' = (\sigma\sqrt{2}\, t)'$ soit $dx = \sigma\sqrt{2}\, dt$

Alors :

$$\text{Var}(x) = \int_{-\infty}^{+\infty} t^2 2\sigma^2 \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-t^2\right) \sigma\sqrt{2}\, dt$$

$$\text{Var}(x) = \frac{\sigma^2}{\sqrt{\pi}} \int_{-\infty}^{+\infty} t.2t \exp\left(-t^2\right) dt$$

$$\text{Var}(x) = \frac{\sigma^2}{\sqrt{\pi}} \left[\left[-t \exp\left(-t^2\right) \right]_{-\infty}^{+\infty} + \int_{-\infty}^{+\infty} \exp\left(-t^2\right) dt \right]$$

or $\left[-t \exp\left(-t^2\right) \right]_{-\infty}^{+\infty} \to 0$ et $\int_{-\infty}^{+\infty} \exp\left(-t^2\right) dt = \sqrt{\pi}$

donc $\text{Var}(x) = \sigma^2$

Ainsi les moyenne et variance de la loi normale sont respectivement :

$$\boxed{\begin{array}{c} E(x) = \mu \\ \text{Var}(x) = \sigma^2 \end{array}}$$

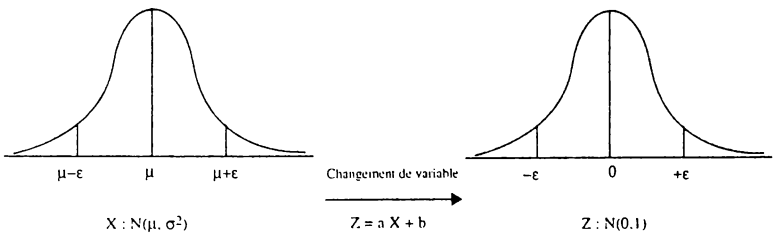
4.3.2. Utilisation de la table de la loi normale N(0,1)

Soit X une variable aléatoire distribuée selon une loi normale quelconque $N(\mu, \sigma^2)$. Pour déterminer la probabilité pour que la variable X soit comprise dans l'intervalle [a ; b], on peut effectuer le calcul suivant :

$$P(a \le x \le b) = \int_a^b N(x)\,dx$$

soit $P(a \le x \le b) = \int_a^b \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right) dx$

Mais pour éviter des calculs longs et fastidieux, ***on utilise la table de la loi normale qui permet de déterminer le résultat rapidement. Il s'agit de la table de la loi normale centrée réduite N(0, 1), c'est-à-dire de moyenne 0 et d'écart-type 1.*** Or les données observées distribuées normalement ne sont pas systématiquement distribuées selon une loi normale centrée réduite. C'est pourquoi **on est souvent amené à effectuer un changement de variable pour passer d'une loi normale quelconque N(m, σ^2) à une loi normale centrée réduite N(0, 1)** afin de pouvoir lire dans la table de la loi normale :



Si la variable aléatoire X tend vers une loi normale N(μ,σ^2), alors après un changement de variable approprié, la variable aléatoire Z = aX + b tendra vers la loi normale centrée réduite N(0,1).

D'après les opérations sur les variables aléatoires :

Si $Z = aX + b$

$$\text{alors} \quad \begin{cases} E(Z) = E(aX+b) \\ Var(Z) = Var(aX+b) \end{cases} \Leftrightarrow \begin{cases} E(Z) = aE(X)+b \\ Var(Z) = a^2\,Var(X) \end{cases} \qquad (1)$$

On veut obtenir une variable aléatoire Z distribuée selon une loi normale centrée réduite N(0,1), c'est-à-dire de moyenne 0 et de variance 1, alors :

$$\text{on connaît} \quad \begin{cases} E(Z) = 0 \\ Var(Z) = 1 \end{cases} \text{ puis on sait que } \begin{cases} E(X) = \mu \\ Var(X) = \sigma^2 \end{cases}$$

$$\text{donc} \quad (1) \Leftrightarrow \begin{cases} 0 = a\mu + b \\ 1 = a^2\sigma^2 \end{cases} \Rightarrow a^2 = \frac{1}{\sigma^2} \Rightarrow \begin{cases} a = \frac{1}{\sigma} \\ b = -\frac{\mu}{\sigma} \end{cases}$$

d'où

$$Z = \frac{1}{\sigma}X - \frac{\mu}{\sigma} = \frac{X-\mu}{\sigma}$$

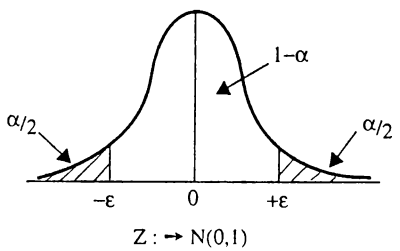
Ainsi pour passer d'une loi normale quelconque N(μ, σ^2) à une loi normale centrée réduite N(0, 1) on effectue le changement de variable suivant :

$$X : N(\mu,\sigma^2) \xrightarrow[Z = \frac{X - \mu}{\sigma}]{\text{Changement de variable}} Z : N(0,1)$$

Alors que pour le changement de variable la notation anglaise est z, en français on note plutôt :

$$\boxed{\varepsilon = \frac{X - \mu}{\sigma}}$$

La table de la loi normale (voir table 1 en annexe) donne la valeur de la probabilité α correspondant à la surface sous la courbe de la loi normale centrée réduite N(0, 1) extérieure à l'intervalle $[-\varepsilon ; +\varepsilon]$, c'est-à-dire la surface hachurée de la figure suivante, soit : $P(Z > \varepsilon \text{ ou } Z < -\varepsilon) = \alpha$



La probabilité complémentaire de la probabilité α est $1 - \alpha$ et correspond donc à la probabilité intérieure à l'intervalle $[-\varepsilon ; +\varepsilon]$, c'est-à-dire :

$$\boxed{P(-\varepsilon < Z < +\varepsilon) = 1 - \alpha}$$

La loi normale centrée réduite étant symétrique autour de 0, on a :

$$\boxed{P(Z > 0) = P(Z < 0) = 1/2 = 50\ \%}$$

puis, par symétrie on a :

$$\boxed{P(Z > \varepsilon) = P(Z < -\varepsilon) = \frac{\alpha}{2}}$$

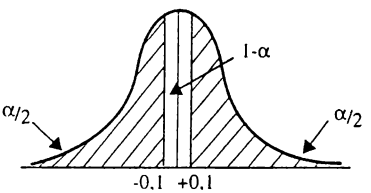
La table de la loi normale centrée réduite donne le risque α correspondant à ε et *vice versa*. Par exemple, à la valeur $\varepsilon = 0{,}1$ correspond la probabilité $\alpha = 0{,}92 = 92\ \%$, c'est-à-dire que si Z est distribuée selon une loi normale centrée réduite N(0,1), alors :

$$P(Z > 0,1 \text{ ou } Z < -0,1) = \alpha = 0,92 = 92\ \%$$

et $$P(Z > 0,1) = P(Z < -0,1) = \frac{\alpha}{2} = \frac{0,92}{2} = 0,46 = 46\ \%$$

De plus, $P(-0,1 < Z < +0,1) = 1-\alpha = 1 - 0,92 = 0,08 = 8\ \%$

Ces probabilités peuvent être représentées ainsi :



■ ***Remarques :***

Il est vivement recommandé de représenter graphiquement les probabilités recherchées, avant tout calcul et utilisation de la table de la loi normale, afin d'éviter toute erreur et de favoriser une meilleure compréhension. Notons que si la valeur recherchée n'apparaît pas dans la table, on prend la valeur la plus proche. Enfin, notons que $P(a < Z < b) = P(a \leq Z \leq b)$ car la loi est continue. Par ailleurs, les logiciels statistiques donnent les valeurs précises des probabilités.

4.3.3. Applications

• Exemple :

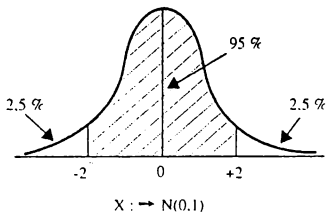
La variable aléatoire X suivant une loi normale centrée réduite N(0,1), calculer :

1. $P(-2 < X < +2)$
2. $P(X > 2)$
3. $P(-1 < X < 2)$
4. $P(X < -1)$
5. $P(X < 1)$
6. X' étant une variable aléatoire normale de moyenne $\mu = 5$ et d'écart-type $\sigma = 2$, calculer : $P(4 < X' < 7)$.

• Correction :

1. $P(-2 < X < +2) = 1 - \alpha$

La valeur recherchée correspond à la zone hachurée.



La valeur la plus proche de 2 dans la table de la loi normale est $|\varepsilon| = 1,96$ et le risque α correspondant est $\alpha = 0,05$

donc $P(-2 < X < +2) = 1 - 0,05$

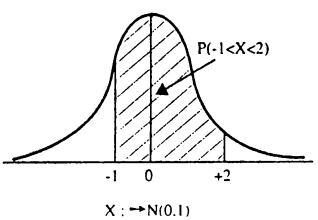
$P(-2 < X < +2) = 0,95 = 95\ \%$

2. $P(X > 2) = \alpha/2$

$P(X > 2) = 0,05/2 = 0,025$

soit $P(X > 2) = 2,5\ \%$

3. La probabilité $P(-1 < X < 2)$ peut être représentée graphiquement par la zone hachurée de la figure suivante :



La table de la loi normale indiquant les probabilités extérieures, on passe alors par la probabilité complémentaire.

$P(-1 < X < 2) = 1 - [P(X > 2) + P(X < -1)]$

$P(-1 < X < 2) = 1 - P(X > 2) - P(X < -1)$

$P(-1 < X < 2) = 1 - P(X > 2) - P(X > +1) = 1 - 0,025 - P(X > +1)$

Or dans la table de la loi normale, la valeur la plus proche de 1 est $\varepsilon = 0,994$ qui correspond à $\alpha = 0,32$, donc $P(X>1) = \alpha/2 = 0,16$.

Alors $P(-1 < X < 2) = 1 - 0,025 - 0,16 = 0,815$

4. $P(X < -1) = P(X > 1) = 0,16$

5. $P(X < 1) = 1 - P(X > 1) = 1 - 0,16 = 0,84$

6. La variable X' suit une loi normale $N(5, 2^2)$, alors pour calculer la probabilité $P(4 < X' < 7)$ on fait un changement de variable $Z = \frac{X'-\mu}{\sigma}$:

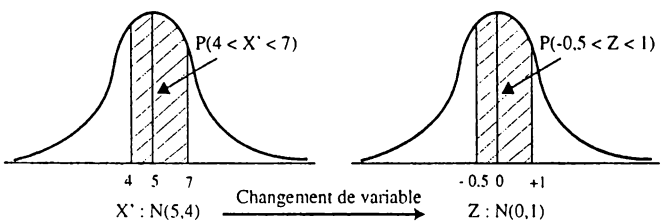
$$P(4 < X' < 7) = P\left(\frac{4-\mu}{\sigma} < \frac{X'-\mu}{\sigma} < \frac{7-\mu}{\sigma}\right)$$

$$= P\left(\frac{4-5}{2} < Z < \frac{7-5}{2}\right)$$

$$= P(-0,5 < Z < 1)$$

Donc $\varepsilon_1 = -0,5$ et $\varepsilon_2 = 1$

Cette probabilité peut être représentée par la zone hachurée de la figure suivante :



On peut donc passer par la probabilité complémentaire :

$P(4 < X' < 7) = P(-0,5 < Z < 1) = 1 - [P(X < -0,5) + P(X > 1)]$

$= 1 - [P(X > 0,5) + P(X > 1)]$

$= 1 - (0,31 + 0,16)$

$= 1 - 0,47 = 0,53 = 53\ \%$

5. Notion d'intervalle de confiance

Dans le paragraphe 2 de ce chapitre, nous avons montré que l'on pouvait caractériser un échantillon i par sa distribution avec une moyenne m_i et une variance s_i^2. Or, chaque fois que l'on tire au sort un échantillon i avec une distribution définie par les paramètres (m_i, s_i^2), d'une population de référence, on cherche à approcher au mieux les caractéristiques de la distribution de paramètres (μ, σ^2) de cette population.

Lorsque l'on souhaite évaluer, dans une population, la valeur moyenne d'un caractère quantitatif (le poids, l'âge ou la taille), on ne dispose le plus souvent que d'un échantillon extrait de cette population. Même lorsque ces

échantillons sont représentatifs, obtenus par tirage au sort, la moyenne observée s'écarte plus ou moins de la moyenne vraie de la population selon le hasard de l'échantillonnage.

Le problème se pose de savoir **dans quelle mesure l'échantillon peut renseigner de manière fiable sur la population de référence**. Quand il n'est pas possible de connaître exactement un paramètre d'une population, les statistiques permettent de déterminer les limites dans lesquelles devrait se trouver ce paramètre (théorique), avec un risque d'erreur noté α. Ces limites constituent ce que l'on appelle ***l'intervalle de confiance du paramètre***. Nous en évoquerons deux ici, l'intervalle de confiance d'une moyenne et celui d'un pourcentage.

Pour un caractère quantitatif on se propose de décrire les **fluctuations de la moyenne** observée m_i autour de la moyenne théorique μ en évaluant, pour chaque intervalle centré sur μ, la probabilité pour que m_i sorte de cet intervalle. Il en est de même pour un **caractère qualitatif**, pour lequel on se propose de décrire les **fluctuations de la proportion p** autour de la proportion théorique π en évaluant, pour chaque intervalle centré sur π, la probabilité pour que p sorte de cet intervalle.

Par ailleurs, lorsque l'on connaît la valeur des paramètres théoriques d'une population de référence, on peut vouloir déterminer les valeurs possibles prises par les statistiques d'un quelconque échantillon issu de cette population. **L'intervalle de pari autour d'une moyenne ou d'une proportion observée** permet d'évaluer les valeurs possibles de ces statistiques. L'intérêt de la recherche de l'intervalle de pari autour des statistiques est notamment de confirmer qu'un échantillon est bien représentatif d'une population de référence.

Enfin, plus la taille d'un échantillon est grande, plus l'intervalle de pari autour d'une statistique (moyenne ou proportion) est petit, et si la statistique observée de l'échantillon est comprise dans cet intervalle de pari, alors plus l'échantillon est représentatif de la population de référence.

5.1. Variance d'une moyenne

Le plus souvent, au lieu de rechercher la valeur exacte de la moyenne théorique μ par l'examen de toute la population, on estime μ à partir d'un ou plusieurs échantillons représentatifs, en consentant à l'avance un certain **risque d'erreur α, lié aux fluctuations d'échantillonnage**.

Si l'on tire au sort N échantillons à partir d'une population de référence, alors on obtient N moyennes (m_1, m_2, ..., m_N), soit une distribution des moyennes m_i de chaque échantillon de moyenne μ. Ces moyennes m_i ne s'écarteront pas sensiblement de μ ; beaucoup seront proches et d'autres

s'en écarteront un peu plus. En fait, si la taille des échantillons est au moins de 30 ($N \geq 30$), alors la distribution des moyennes sera normale de moyenne μ.

Il faut bien distinguer cette variance de la moyenne (qui dépend de la taille n de l'échantillon) de celle s^2 de l'échantillon. En supposant que chacun des N échantillons soit de taille n, cette distribution sera normale avec une variance σ_m^2 autour de cette moyenne qui vaut :

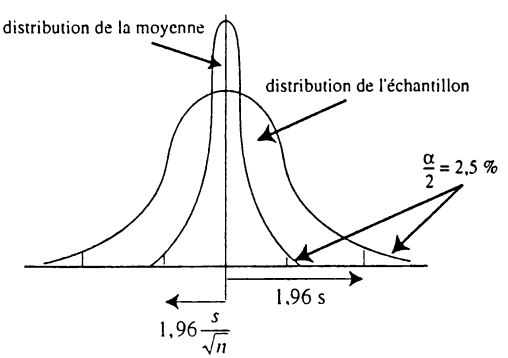
Variance de la moyenne : $\boxed{\sigma_m^2 = \frac{\sigma^2}{n}}$

On peut estimer cette variance sur un échantillon avec une bonne précision en utilisant la variance s^2 de l'échantillon quand il est suffisamment grand ($n \geq 30$) :

Variance de l'échantillon ($n \geq 30$) : $\boxed{s_m^2 = \frac{s^2}{n}}$

Pour bien distinguer la distribution de l'échantillon de celle de la moyenne, nous les représentons sur le schéma ci-dessous. La distribution de l'échantillon peut être figurée par une loi normale $N(m ; s^2)$, alors que la distribution de la moyenne autour de μ est une loi normale $N(\mu ; \frac{\sigma^2}{n})$ que l'on a approximée par $N(m ; \frac{s^2}{n})$.

Autrement dit, un individu tiré au sort dans la population a 95 % de chance d'être inclus dans l'intervalle $[\mu - \varepsilon_{5\%} s ; \mu + \varepsilon_{5\%} s]$ tandis qu'un échantillon a 95 % de chance d'avoir une moyenne incluse dans l'intervalle $\left[\mu - \varepsilon_{5\%} \frac{s}{\sqrt{n}} ; \mu + \varepsilon_{5\%} \frac{s}{\sqrt{n}}\right]$. On note $\varepsilon_{5\,\%}$, ε correspondant au seuil $\alpha = 5$ %.



Si l'on tire 100 échantillons dans une population, alors la probabilité que chaque moyenne tombe dans l'intervalle $\left[\mu - 1{,}96\frac{s}{\sqrt{n}} ; \mu + 1{,}96\frac{s}{\sqrt{n}}\right]$ est de 0,95 (puisque $\varepsilon_{5\,\%}=1{,}96$).

Autrement dit, sur ces 100 échantillons, on s'attend à ce que 5 d'entre eux soient à l'extérieur de l'intervalle : on accepte ce risque comme suffisamment faible, c'est pourquoi il est traditionnellement choisi.

5.2. Intervalle de pari d'une moyenne

Supposons que les valeurs théoriques μ et σ^2 soient connues. Dans ce cas, nous savons que, pour un échantillon de n sujets ($n \geq 30$), on peut considérer que la moyenne observée, m, suit une loi normale de moyenne μ et d'écart-type $\frac{\sigma}{\sqrt{n}}$.

En utilisant l'écart-type $\frac{s}{\sqrt{n}}$, on peut donc admettre avec un risque d'erreur $\alpha = 5\,\%$ (pour lequel $\varepsilon_{5\%}=1{,}96$), que la moyenne observée, m, doit être comprise dans l'intervalle de pari à 95 %, entre :

$$\boxed{\mu - 1{,}96\ \frac{s}{\sqrt{n}}\ ;\ \mu + 1{,}96\ \frac{s}{\sqrt{n}}}$$

avec $1{,}96 = \varepsilon$, correspondant au risque $\alpha = 5\,\%$, lu dans la table de la loi normale (table 1).

On peut encore exprimer l'intervalle de pari en disant que la probabilité pour que la moyenne d'un échantillon soit comprise dans cet intervalle est de 95 % :

$$P\left(\mu - 1{,}96\frac{s}{\sqrt{n}} \leq m \leq \mu + 1{,}96\frac{s}{\sqrt{n}}\right) = 95\%$$

■ ***Notations :***

Pour présenter l'intervalle de pari à 95 % limité par les bornes constantes quelconques a et b, on note $IP_{95\,\%}$:[a,b].

• Exemple :

Supposons que nous connaissions la moyenne d'âge $\mu = 40$ et l'écart-type $\sigma = 5$ d'une population. Nous pouvons, par exemple, constituer un échantillon de taille $n = 100$ personnes et chercher la probabilité pour que la moyenne observée sur cet échantillon soit comprise entre 39 et 41 ans. On recherche donc la probabilité pour que :

$$40 - 1 \leq m \leq 40 + 1$$

d'après la formule suivante :

$$\mu - 1{,}96\frac{s}{\sqrt{n}} \leq m \leq \mu + 1{,}96\frac{s}{\sqrt{n}}$$

on aboutit à la relation :

$$\varepsilon_\alpha \times \frac{\sigma}{\sqrt{n}} = 1$$

ainsi,

$$\varepsilon_\alpha = 1 \times \frac{\sqrt{n}}{\sigma} = \frac{\sqrt{100}}{5} = 2$$

La table de la loi normale nous donne pour tout écart-réduit ε le risque α correspondant. La valeur la plus proche est $\varepsilon = 1{,}96$, soit un risque $\alpha = 5$ %. On a donc 5 chances sur 100 pour que la moyenne de l'échantillon de 100 personnes sorte de l'intervalle [39 ; 41].

On peut considérer ce risque $\alpha = 0{,}05$ comme faible et poser que, si la moyenne de l'échantillon de 100 personnes n'est pas dans l'intervalle, alors celui-ci n'est pas représentatif de la population.

Nous pouvons aussi réduire l'intervalle de pari grâce à un échantillon plus grand. Si l'échantillon est constitué de 2 500 personnes, alors l'intervalle de pari pour un risque $\alpha = 0{,}05$ est :

$$[40 - 1{,}96\frac{5}{50} ; 40 + 1{,}96\frac{5}{50}],$$

soit à peu près

$$[39{,}8 ; 40{,}2].$$

5.3. Intervalle de confiance d'une moyenne

5.3.1. Cas de grands échantillons, $n \geq 30$:

Pour estimer μ à partir d'un échantillon, on peut dire que l'observation d'une moyenne m sur un échantillon de n observations permet d'assigner à la moyenne théorique inconnue μ, l'intervalle de confiance à 95 % (soit 1-α) :

$$\boxed{\left[m - \varepsilon_{5\%}\frac{s}{\sqrt{n}} ; m + \varepsilon_{5\%}\frac{s}{\sqrt{n}}\right]}$$

La formule reflète bien le fait que **plus l'échantillon est grand, plus l'on se rapproche de la moyenne théorique**.

On peut l'exprimer autrement en disant que la probabilité pour que la moyenne théorique μ de la population de référence d'un échantillon soit comprise dans cet intervalle de confiance est de 95 % :

$$P\left(m - 1{,}96\frac{s}{\sqrt{n}} \leq \mu \leq m + 1{,}96\frac{s}{\sqrt{n}}\right) = 95\ \%$$

▩ *Conditions d'application :*
$n \geq 30$.

▩ *Notation :*
Pour présenter l'intervalle de confiance à 95 % limité par les bornes constantes quelconques a et b, on note $IC_{95\ \%}$:[a,b].

5.3.2 Cas de petits échantillons n < 30

L'observation d'une moyenne m sur un échantillon de n cas permet d'assigner à la moyenne inconnue μ l'intervalle de confiance à 95 % :

$$\left[m - t_{5\%}\ \frac{s}{\sqrt{n}}\ ;\ m + t_{5\%}\ \frac{s}{\sqrt{n}} \right]$$

avec s : écart-type estimé sur l'échantillon,
t : lu dans la table de t au risque $\alpha = 5\ \%$ avec un nombre de degré de liberté n - 1.

Ainsi, $\mu \in \left[m - t\frac{s}{\sqrt{n}}\ ;\ m + t\frac{s}{\sqrt{n}}\right]$.

▩ *Conditions d'application :*
La distribution de la variable étudiée doit être normale.

5.4. Intervalle de pari d'un pourcentage

Si l'on fait des hypothèses sur la valeur du pourcentage π, il est possible de calculer le risque α pour que le pourcentage d'un échantillon sorte de l'intervalle :

$$\left[\pi - \sqrt{\frac{\pi(1-\pi)}{n}}\ ;\ \pi + \sqrt{\frac{\pi(1-\pi)}{n}} \right]$$

On peut choisir comme intervalle de pari [40 % ; 60 %] pour une probabilité théorique π et déterminer le risque α qui lui est associé.

▩ *Conditions d'application :*
$n\pi$ et $n(1-\pi) \geq 5$.

• Exemple :

Supposons que l'on demande à 100 personnes d'une commune tirées au hasard leur intention de vote pour deux candidats. Dans ce cas, l'écart-type attendu est :

$$\sigma = \sqrt{\frac{\pi(1-\pi)}{n}}$$

soit, si $\pi = 0{,}5$

$$\sigma = \sqrt{\frac{0{,}5 \times 0{,}5}{100}} = \frac{0{,}5}{10} = 0{,}05$$

La table de l'écart-réduit (table 1) nous donne pour tout écart-réduit ε le risque α correspondant. Avec l'intervalle de pari [0,4 ; 0,6], l'écart autour de 50 % est donc 10 %, soit un écart $e = 0{,}10$ et l'on a :

$$\varepsilon = \frac{e}{\sigma}$$

soit,

$$\varepsilon = \frac{0{,}10}{0{,}05} = 2$$

La valeur la plus proche dans la table est $\alpha = 0{,}05$ pour $\varepsilon = 1{,}96$. Avec un échantillon représentatif de 100 personnes, on a donc 5 chances sur 100 de trouver un pourcentage en dehors de l'intervalle [40 % ; 60 %] quand le pourcentage théorique est $\pi = 0{,}5$.

5.5. Intervalle de confiance d'un pourcentage

De même que pour une moyenne, lorsque l'on veut évaluer dans une population le pourcentage π des sujets qui présentent un caractère donné, on s'appuie sur un échantillon représentatif. Dans ce cas, le pourcentage observé p s'écarte plus ou moins de cette valeur π selon les fluctuations d'échantillonnage.

L'observation d'un pourcentage p sur un échantillon de n cas permet d'assigner au pourcentage inconnu π l'intervalle de confiance à 95 % (95 % correspondant à 1-α), avec $\varepsilon_{5\,\%} = 1{,}96$:

$$p \pm 1{,}96\sqrt{\frac{pq}{n}}$$

soit

$$\boxed{\left[p - 1{,}96\sqrt{\frac{pq}{n}} \; ; \; p + 1{,}96\sqrt{\frac{pq}{n}}\right]}$$

▩ *Conditions d'application :*

np et $nq \geq 5$, avec $q = 1-p$.

6. Exercices

6.1. Exercices corrigés (solutions p. 113 et suivantes)

• Exercice 1 :

A la suite de la fécondation d'un ovule "X" par un spermatozoïde "X" ou "Y", le sexe de l'enfant est déterminé par le fait que le spermatozoïde porte un chromosome sexuel X, on a alors XX : une fille ; ou par le fait que le spermatozoïde porte un chromosome sexuel Y, on a alors XY : un garçon. Les vrais jumeaux résultent d'un dédoublement de l'ovule fécondé : les enfants sont alors de même sexe. Les faux jumeaux résultent de la fécondation par deux spermatozoïdes différents de deux ovules : les enfants peuvent alors être de même sexe ou de sexe différent.

1. Quelles sont les probabilités respectives que deux jumeaux soient : a) deux garçons, b) deux filles, c) une fille et un garçon ; lorsque ce sont des vrais jumeaux ?

2. Même question quand il s'agit de faux jumeaux.

3. Quelles sont les probabilités que l'on ait affaire à de vrais jumeaux ou à de faux jumeaux, selon que l'on a des jumeaux : a) deux garçons, b) deux filles, ou c) un garçon et une fille ?

On posera p la fréquence des vrais jumeaux parmi les jumeaux, et q = 1-p pour les faux jumeaux.

• Exercice 2 :

Lors d'une épidémie d'une maladie M, la prévalence de la maladie M est de 10 %. La spécificité du signe diagnostique S, pour cette maladie, est de 0,4 et sa sensibilité est de 0,9.

En période d'épidémie, quelle est la probabilité qu'un sujet soit atteint de cette maladie lorsque le signe diagnostique est révélé positif ?

• Exercice 3 :

Il a été démontré que, dans certains pays, le pourcentage de personnes dépressives décroît avec le nombre de postes téléphoniques par habitant de la population. Interprétez cette constatation.

• Exercice 4 :

Une enquête en psychologie sociale, portant sur un échantillon de 1 500 personnes, indique que 41 % d'entre elles ont une attitude de leader en situation de groupe.

1. Donner l'intervalle de confiance à 95 % de la proportion théorique des personnes de la population de référence qui ont une attitude de leader.

2. Même question si le pourcentage observé avait été obtenu à partir d'un échantillon de 500 personnes, plutôt qu'à partir d'un échantillon de 1 500 personnes.

6.2. Exercices non corrigés

• *Exercice 5 :*

De combien de manières peut-on former une combinaison de 3 femmes et de 4 hommes choisis parmi 7 femmes et 6 hommes ?

• *Exercice 6 :*

Soit une population composée de 45 % d'hommes. On suppose que 8 % des hommes et 3,3 % des femmes sont atteints d'une pathologie "P".

1. Quelle est la probabilité qu'une personne soit atteinte de cette pathologie "P"?
2. Sachant qu'une personne est atteinte par cette pathologie "P", quelle est la probabilité que ce soit un homme ? En déduire la probabilité que ce soit une femme.

• *Exercice 7 :*

Un test diagnostique de la dépression a une spécificité de 0,7 et une sensibilité de 0,9. Chez un groupe de sujets venant consulter en psychiatrie, 40 % d'entre eux ont réellement une dépression.

1. Combien de sujets seront diagnostiqués dépressifs à tort selon ce test ?
2. Quelle est la valeur prédictive négative du test ?
3. Quelle est la valeur prédictive positive du test ?
4. Que se passe-t-il lorsque l'on augmente la sensibilité du test ?

• *Exercice 8 :*

Lors du lancer de 5 dés, où trois essais sont permis (en conservant au moins 2 dés dont les chiffres obtenus sont identiques), quelle est la probabilité d'obtenir le même chiffre pour les 5 dés ?

• *Exercice 9 :*

Au loto, on tire au sort 6 numéros parmi 49 numéros. De son côté, le joueur parie sur 6 numéros.

1. Quelle est la probabilité d'obtenir exactement 3 numéros gagnants parmi ses 6 numéros ?
2. Quelle est la probabilité d'obtenir exactement 4 numéros gagnants parmi ses 6 numéros ?
3. Quelle est la probabilité d'obtenir exactement 5 numéros gagnants parmi ses 6 numéros ?

4. Quelle est la probabilité d'obtenir exactement 6 numéros gagnants parmi ses 6 numéros ?
5. Au cours d'un tirage particulier, il y a eu 4 joueurs ayant les 6 bons numéros, 2 354 joueurs ayant 5 bons numéros, 96 096 joueurs ayant 4 bons numéros et 2 599 060 joueurs ayant 3 bons numéros. Combien y a-t-il eu de joueurs ? Commenter ce résultat.

• *Exercice 10 :*

On considère deux signes diagnostiques A et B. La présence du signe est notée "+"; l'absence est notée "-". Les sujets "A+" sont donc ceux ayant le signe A et ayant ou non le signe B. Les sujets "A+B-" sont ceux ayant le signe A et n'ayant pas le signe B.
1. Chez les sujets "normaux" les proportions des diverses combinaisons sont les suivantes :

A-B-	A-B+	A+B-	A+B+
1/10	1/5	1/5	1/2

Les signes A et B sont-ils indépendants ?

2. Soit une pathologie P affectant un sujet sur 30, les proportions suivantes sont chez les malades :

A-B-	A-B+	A+B-	A+B+
1/2	1/4	0	1/4

a. On examine un sujet : il a le signe B. Quelle probabilité a-t-il d'être malade ?
b. Si ce sujet présente le signe B et ne présente pas le signe A (il est A-B+). Quelle probabilité a-t-il d'être malade ?

• *Exercice 11 :*

Une étude de recherche en psychologie de la santé a permis d'observer un échantillon de 2 000 personnes hospitalisées dans différents services cliniques ayant consulté au moins une fois un psychologue lors de leur séjour. En moyenne, chaque consultation avec un psychologue a une durée de 20 minutes, avec un écart-type de 30.
1. Donner l'intervalle de confiance à 95 % de la moyenne théorique de la durée des consultations en psychologie dans cet hôpital.
2. Même question si la moyenne observée avait été obtenue à partir d'un échantillon de 800 personnes, plutôt qu'à partir d'un échantillon de 2 000 personnes.

• *Exercice 12 :*

On souhaite faire un essai thérapeutique contrôlé chez les adolescents présentant des migraines, afin de déterminer l'effet placebo *versus* paracétamol. Indiquer la meilleure répartition en un groupe traité et un groupe témoin parmi les solutions proposées ci-dessous, en expliquant clairement chacune de vos décisions :

1. Dans un collège, en ne traitant que les adolescents présentant des migraines d'une seule classe et en prenant les autres comme témoins.

2. En prenant comme référence les résultats obtenus dans une ville, à la suite d'un essai thérapeutique par paracétamol chez les adolescents présentant des migraines, l'année précédant un nouvel essai par placebo dans cette même ville.

3. Dans un collège, en prenant comme témoins les adolescents dont les parents refusent le traitement par paracétamol.

4. Dans deux collèges d'une même ville, où dans l'un on traitera tous les adolescents présentant des migraines par paracétamol et dans l'autre, où tous les adolescents présentant des migraines seront sous placebo.

5. Autre solution - laquelle ?

6.3. Solutions

• *Correction de l'exercice 1 :*

1. Vrais jumeaux (VJ) :

a. P(GG/VJ) = 1/2

b. P(FF/VJ) = 1/2

c. P(FG/VJ) = 0

2. Faux jumeaux (FJ) :

a. P(GG/FJ) = 1/4

b. P(FF/FJ) = 1/4

c. P(FG/FJ) = 1/2

3. On utilise le théorème de Bayes

a.
$$P(VJ / GG) = \frac{P(GG / VJ)\,P(VJ)}{P(GG / VJ)\,P(VJ) + P(GG / FJ)\,P(FJ)}$$

$$P(VJ / GG) = \frac{1/2 \times p}{1/2 \times p + 1/4 \times (1-p)} = \frac{2p}{p+1}$$

$$P(FJ / GG) = \frac{P(GG / FJ)\,P(FJ)}{P(GG / FJ)\,P(FJ) + P(GG / VJ)\,P(VJ)}$$

$$P(FJ / GG) = 1 - P(VJ / GG) = 1 - \frac{2p}{p+1} = \frac{1-p}{1+p}$$

b. $$P(VJ / FF) = \frac{P(FF / VJ)\,P(VJ)}{P(FF / VJ)\,P(VJ) + P(FF / FJ)\,P(FJ)}$$

$$P(VJ / FF) = \frac{1/2 \times p}{1/2 \times p + 1/4 \times (1-p)} = \frac{2p}{p+1}$$

Et ainsi de suite...
On peut résumer les résultats dans le tableau suivant, qui se lit pour la première cellule du tableau : probabilité d'avoir de vrais jumeaux lorsque l'on a deux garçons, soit P(VJ/GG).

Probabilité →	/GG	/FF	/FG
VJ	$\frac{2p}{1+p}$	$\frac{2p}{1+p}$	0
FJ	$\frac{1-p}{1+p}$	$\frac{1-p}{1+p}$	1

• *Correction de l'exercice 2 :*

Soit M : présenter la maladie M,
$\overline{M}$: ne pas présenter la maladie M,
S : diagnostic positif (présence du signe S),
$\overline{S}$: diagnostic négatif (absence du signe S).

$P(M) = 10\,\% = 0{,}1$
$Sp = P(\overline{S}/\overline{M}) = 0{,}4$
$Se = P(S/M) = 0{,}9$
Alors :

$$P(M / S) = \frac{P(S / M)\,P(M)}{P(S / M)\,P(M) + P(S / \overline{M})\,P(\overline{M})}$$

$$P(M / S) = \frac{0{,}9 \times 0{,}1}{0{,}9 \times 0{,}1 + 0{,}6 \times 0{,}9} = \frac{0{,}09}{0{,}63} = 0{,}1429 = 14{,}29\,\%$$

Cette probabilité correspond à la probabilité de présenter la maladie M lorsque le signe diagnostique est positif, il s'agit de la valeur prédictive positive (VPP) du signe diagnostique S. Le tableau suivant présente toutes les probabilités conditionnelles possibles (avec VPN correspondant à la valeur prédictive négative).

Probabilité →	+ Présence du signe diagnostique	- Absence du signe diagnostique
+ Malade	$P(M / S)$ VPP = 14,29 %	$P(M / \overline{S})$
- Non malade	$P(\overline{M} / S)$	$P(\overline{M} / \overline{S})$ VPN

• *Correction de l'exercice 3 :*

Pour considérer que cette constatation peut conduire à une conclusion de cause à effet, c'est-à-dire que le nombre de postes de téléphone serait un facteur prédictif de la prévalence de la dépression, ou serait un moyen de lutte contre la dépression, il faudrait alors s'assurer que toutes choses sont égales par ailleurs. Or le nombre de postes téléphoniques pourrait être un facteur de confusion (ou facteur confondant) du facteur socio-économique (le nombre de postes pouvant être lié au statut), du facteur situation géographique (le nombre de postes pouvant dépendre du degré d'isolement, de la densité de population), ou du facteur chef de famille ou non (l'abonnement au téléphone ne concernant qu'une ou deux personnes par famille, cela limite alors l'exploration des dépressions des jeunes vivant chez leurs parents et des sujets n'étant pas directement abonnés au téléphone).

Ainsi, pour envisager de conclure à une relation de cause à effet, il est essentiel de procéder au préalable lors de la constitution des échantillons à un tirage au sort afin d'obtenir un échantillon représentatif de la population, en s'assurant que les sujets sont comparables en dehors des variables étudiées, c'est-à-dire que toutes choses sont égales par ailleurs. Notons qu'une étude de corrélation entre la variable nombre de postes téléphoniques et la variable prévalence de la dépression, ne conduirait pas non plus à une relation de cause à effet, mais simplement à une association entre les deux variables étudiées. De plus, il faudrait vérifier que la cause survient avant l'effet. Ici, cela reviendrait à vérifier que la dépression survient après l'acquisition de plusieurs postes téléphoniques !

• *Correction de l'exercice 4 :*

1. On connaît les paramètres suivants :
$p_0 = 0{,}41$; $q_0 = 1 - p_0 = 0{,}59$; $N = 1\,500$.
Les conditions d'application de l'intervalle de confiance autour d'une proportion sont vérifiées, puisque $N \times p_0$ et $N \times q_0$ sont ≥ 5.
La lecture du paramètre tabulé ε_T correspondant à un risque α de 5 %, lu dans la table de la loi normale, est équivalent à 1,96.

$$IC_{(95\,\%)} = \left[p_0 - \varepsilon_{5\,\%} \sqrt{\frac{p_0 q_0}{N}} \;\; ; \;\; p_0 + \varepsilon_{5\,\%} \sqrt{\frac{p_0 q_0}{N}} \right]$$

$$IC_{(95\,\%)} = \left[0{,}41 - 1{,}96 \sqrt{\frac{0{,}41 \times 0{,}59}{1500}} \;\; ; \;\; 0{,}41 + 1{,}96 \sqrt{\frac{0{,}41 \times 0{,}59}{1500}} \right]$$

$$IC_{(95\,\%)} = [0{,}385 \;\; ; \;\; 0{,}435]$$

2. Avec $N = 500$,
$IC_{(95\,\%)} = [0{,}367 \;\; ; \;\; 0{,}453]$

Plus la taille de l'échantillon est grande, plus l'intervalle de confiance est précis autour d'une proportion (donc plus l'intervalle est petit). Notons que cela est également le cas autour d'une moyenne.

CHAPITRE IV

Statistiques descriptives : l'analyse préliminaire

1. Description des données *119*
1.1. Description de la population étudiée *119*
1.2. Description des variables *120*
1.3. Regroupement des données par classes *122*
1.4. Tri à plat des données *123*
1.4.1. Rôle *123*
1.4.2. Procédure *123*

2. Représentations graphiques des données *125*
2.1. Diagrammes en bâtons *125*
2.1.1. Rôle *125*
2.1.2. Procédure *125*
2.2. Histogrammes *127*
2.2.1. Rôle *127*
2.2.2. Procédure *127*
2.3. Camemberts *129*
2.3.1. Rôle *129*
2.3.2. Procédure *129*

CHAPITRE IV

Statistiques descriptives : l'analyse préliminaire

1. Description des données

1.1. Description de la population étudiée

Le traitement statistique descriptif des données ne permet de représenter qu'une partie de l'information issue des observations initiales. Cependant, cette perte d'information inévitable permet de synthétiser et de résumer les observations de différentes manières, et permet d'avoir une vision d'ensemble la plus claire possible, que cela soit à travers les tableaux ou les graphiques (histogramme, diagramme en bâtons, camembert, courbe).

L'étude de la distribution d'une variable déterminera l'homogénéité des données ou au contraire mettra en évidence une bimodalité, voire une plurimodalité des observations. Une bimodalité ou une plurimodalité des données permet de visualiser directement deux, voire plusieurs, **sous-groupes de sujets** répondant davantage à certaines modalités d'un caractère. Dans ce cas, l'utilisation de la moyenne pour représenter les données ne sera pas informative et pourra conduire à de fausses interprétations des résultats. Par exemple, si la distribution des notes à un examen présente deux modalités, l'une autour de 18 et l'autre autour de 3, et que la moyenne de l'ensemble des notes est de 10. Dans ce cas, le simple résumé des observations par la moyenne laissera penser que les sujets ont dans l'ensemble réussi moyennement leur examen, alors qu'en réalité deux catégories de sujets se distinguent, ceux qui ont très bien réussi leur examen et ceux qui ne l'ont pas réussi.

Ainsi, une fois les données recueillies, avant de se lancer vers une analyse des données en terme de vérification d'hypothèses préalables, il est essentiel, dans un premier temps, d'examiner attentivement les données, de représenter graphiquement leur distribution et les séries de moyennes observées. Cette démarche a également l'intérêt de faire parfois apparaître visuellement des différences entre les sous-groupes de sujets, et permet alors de poser de nouvelles hypothèses à tester. Elle permet aussi d'envisager des regroupements des données par classe et la création de nouvelles variables.

De plus, l'étude de la forme des distributions, notamment la recherche de la conformité des distributions observées au modèle de la loi normale, permettra de vérifier les conditions d'application requises par les tests statistiques envisagés dans la suite de l'analyse des données.

Les statistiques descriptives constituent donc une étape préliminaire essentielle à l'analyse des données proprement dite.

1.2. Description des variables

La description des variables consiste en l'établissement de la liste exhaustive des variables qualitatives et quantitatives, et de leurs modalités de réponses spécifiques. A cette étape, l'utilisation de tableurs (logiciel informatique tel que Excel), ou un comptage à la main dans le cas de faibles effectifs, permet pour chaque variable qualitative de dénombrer les occurrences de chacune de ses modalités de réponse ; puis pour chaque variable quantitative de déterminer son étendue, c'est-à-dire son intervalle de variation (les valeurs minimum et maximum observées) et de déterminer le nombre d'observations de même valeur.

Par exemple, pour la variable qualitative "sexe" on peut dénombrer les femmes et les hommes (nombre de femmes et nombre d'hommes). Puis, pour la variable quantitative "âge" on peut procéder à des regroupements par classe pour dénombrer par exemple le nombre de sujets de moins de 30 ans et le nombre de sujets de 30 ans et plus ; ou encore déterminer l'étendue observée de cette variable, par exemple 7 à 77 ans.

Chaque variable quantitative peut être résumée dans des tableaux par ses caractéristiques numériques majeures (voir chapitre III), telles que sa moyenne, son étendue, son écart-type, et éventuellement sa médiane, son mode, son écart-moyen et sa variance (voir tableaux 1 et 2 suivants).

• *Exemple :*

On étudie l'évolution d'une pathologie P à deux stades de gravité (I et II) chez 100 patients traités, suivis pendant 6 mois. Pour chaque patient, on surveille les effets secondaires liés au traitement, et on mesure un score d'une échelle psychométrique X une fois par mois. Les tableaux 1 et 2 suivants résument les données observées.

Age (années)	m ± s	47,9 ± 10,4
Nombre total de patients		100
Sexe n (%)	féminin	100 (100 %)
Taille (cm)	m ± s	159,3 ± 7,6
Poids (kg)	m ± s	65,5 ± 9,1
Score observé (échelle psychométrique)	m ± s	16,7 ± 10
Score sur les 6 mois	étendue	0 - 60
Score par classe n (%)	0 à 29	90 (90 %)
	30 à 60	10 (10 %)

m : moyenne
s : écart-type

Tableau 1

Numéro des mesures (mois)	1	2	3	4	5	6
Nombre de patients	100	100	100	98	97	97
Stade I n (%)	80 (80,0)	80 (80,0)	80 (80,0)	70 (71,4)	67 (69)	50 (51,5)
Stade II n (%)	20 (20,0)	20 (20,0)	20 (20,0)	28 (28,6)	30 (31)	47 (48,5)
Nombre de sujets sortis d'étude	0	0	0	2	3	3
Nombre de sujets avec effets secondaires	6	5	7	5	0	6
Durée moyenne des effets secondaires en jours (s)	2,3 (1,5)	3,2 (3,2)	2,0 (1,7)	3,0 (4,0)	- (-)	4,0 (4,4)
Etendue de la durée	1-4	1-8	1-4	1-9	-	1-9

Tableau 2

1.3. Regroupement des données par classes

Le regroupement des données par classes est utilisé dans le cas de variables d'intervalle à multiples valeurs, telles que les variables âge, taille ou poids des sujets.

Les classes doivent être formées de telle sorte que :

– **Chaque observation appartient à une classe** ; c'est-à-dire que les classes doivent être exhaustives afin de comprendre toutes les observations possibles.

– **Chaque observation appartient à une seule classe.** On parle de classes incompatibles, où chaque limite d'intervalle est différente d'une classe à l'autre (voir la variable "score" classée dans le tableau 1 ci-dessus, les observations d'un score égal à 30 ne se retrouvent que dans une seule classe, ici la deuxième).

– **Chaque classe doit avoir la même amplitude** (intervalles de même taille), afin de pouvoir calculer des indices numériques et éventuellement d'attribuer à chaque classe sa valeur centrale. Les classes extrêmes peuvent cependant échapper à cette règle. Par exemple, lors des classes d'âge, la première classe peut comprendre tous les sujets de moins de 15 ans, et la dernière classe tous ceux de plus de 65 ans. Par ailleurs, on peut aussi former les bornes des classes en fonction des effectifs, par exemple en constituant des classes comprenant chacune 10 % des sujets observés. Dans ce dernier cas les classes n'auront pas la même amplitude.

– **Le nombre de classes ne doit pas être trop élevé** (autour de 10 maximum en général), afin d'obtenir une présentation des données la plus claire possible et de faciliter les calculs.

▩ *Remarques :*

Dans le cas de variables qualitatives, on peut être amené à regrouper plusieurs classes préalablement constituées au début de l'étude, si les effectifs sont faibles dans certaines classes.

De plus, on peut être amené à transformer certaines variables quantitatives en variables qualitatives en formant plusieurs classes. On parle alors de variables quantitatives classées ou semi-quantitatives.

L'objectif du regroupement des données par classe facilite les calculs et résume les observations. Mais les moyens informatiques disponibles actuellement permettent d'éviter ce regroupement en favorisant une plus grande précision des résultats.

1.4. Tri à plat des données

1.4.1. Rôle

L'objectif du tri à plat est de représenter les données observées sous forme de **tableaux**, afin de résumer les effectifs et répartitions de toutes les modalités de chaque variable qualitative ou quantitative classée.

1.4.2. Procédure

Pour chaque variable on forme un tableau (voir tableaux 3 et 4), dans lequel figure :

– **à la première ligne**, l'intitulé de toutes les modalités de la variable ;

– **à la deuxième ligne**, le nombre d'observations de chaque modalité de la variable, c'est-à-dire l'effectif de chaque modalité, constituant ainsi la distribution observée de la variable appelée tri à plat (encore appelée distribution expérimentale, distribution statistique ou série statistique) ;

– **à la troisième ligne**, la fréquence (ou proportion) des observations de chaque modalité de la variable, c'est-à-dire chaque effectif divisé par le nombre total d'observations ;

– **à la quatrième ligne**, éventuellement les pourcentages observés (équivalents aux fréquences observées, hormis le calcul), un pourcentage de 10 % correspondant à une fréquence de 10 divisé par 100 soit 0,10 ;

– **à la cinquième ligne**, les effectifs cumulés croissants, c'est-à-dire que pour chaque modalité on additionne l'effectif de cette modalité à l'ensemble des effectifs des modalités inférieures, cela uniquement dans le cas d'échelles ordinales ou d'intervalles ;

– **à la sixième ligne**, éventuellement les fréquences cumulées croissantes, c'est-à-dire que pour chaque modalité on calcule les fréquences des effectifs cumulés, toujours uniquement dans le cas d'échelles ordinales ou d'intervalles ;

– **à la septième ligne**, éventuellement les pourcentages cumulés croissants, c'est-à-dire que pour chaque modalité on calcule les pourcentages des effectifs cumulés ou on transforme les fréquences cumulées en pourcentages cumulés, toujours uniquement dans le cas d'échelles ordinales ou d'intervalles.

▩ *Remarques :*

Toutes les classes ou modalités possibles doivent être représentées dans le tri à plat, même si elles ne comprennent aucune observation, notamment pour faire apparaître directement les "trous" d'une distribution (c'est-à-dire les valeurs nulles).

Le cumul des effectifs permet notamment de déterminer directement l'effectif d'observations inférieures ou égales à une modalité donnée.

• Exemple 1 :

Lors d'une étude en psychologie sociale sur la mobilité géographique, on a interrogé 50 personnes pour savoir si elles passaient leurs vacances à l'étranger. Puis, on a effectué un tri à plat des données observées, représenté dans le tableau 3.

Nom de la variable : Vacances à l'étranger	Jamais	Parfois	Souvent	Toujours	Total
Effectifs	8	17	25	0	50
Fréquences	0,16	0,34	0,50	0	1
Pourcentage	16 %	34 %	50 %	0	100 %
Effectifs cumulés croissants	8	25	50	50	50
Fréquences cumulées croissantes	0,16	0,50	1	1	1
Pourcentages cumulés croissants	16 %	50 %	100 %	100 %	100 %

Tableau 3

• Exemple 2 :

On a interrogé 200 personnes sur leur âge et on a noté leur réponse par classe d'âge. Puis, on a effectué un tri à plat des données observées, représenté dans le tableau 4.

Nom de la variable : Age	moins de 25 ans	25 à 34 ans	35 à 44 ans	45 à 54 ans	55 à 64 ans	65 ans et plus	Total
Effectifs	20	44	32	50	40	14	200
Fréquences	0,10	0,22	0,16	0,25	0,20	0,07	1
Pourcentage	10 %	22 %	16 %	25 %	20 %	7 %	100 %
Effectifs cumulés croissants	20	64	96	146	186	200	200
Fréquences cumulées croissantes	0,10	0,32	0,48	0,73	0,93	1	1
Pourcentages cumulés croissants	10 %	32 %	48 %	73 %	93 %	100 %	100 %

Tableau 4

2. Représentations graphiques des données

Les données observées peuvent être représentées dans des tableaux, comme nous l'avons vu dans le paragraphe 1., lors des tris à plat ou des tableaux résumés des valeurs caractéristiques des modalités des variables. Or, les représentations graphiques des distributions ont l'avantage de **mettre en évidence plus facilement et plus rapidement les irrégularités ou les régularités des observations**. Elles permettent également de faire apparaître des similitudes ou non entre plusieurs groupes d'individus étudiés, ou éventuellement une ressemblance avec des distributions théoriques connues, telles que la forme en cloche de la loi normale.

Pour que cette lecture rapide soit la plus efficace et la plus précise possible, il est essentiel, non seulement d'apporter toutes les informations nécessaires directement sur les graphiques eux-mêmes afin de ne pas avoir besoin pour les comprendre de se reporter aux commentaires associés, mais aussi de respecter certaines règles selon le type de représentation afin d'éviter des erreurs d'interprétation (par exemple concernant les échelles). Dans cet objectif, le titre du graphique, l'intitulé des variables et de leurs modalités doivent être explicites et lisibles.

2.1. Diagrammes en bâtons

2.1.1. Rôle

L'objectif des diagrammes en bâtons est de visualiser les effectifs de toutes les modalités d'une variable.

2.1.2. Procédure

On reporte sur une ligne horizontale toutes les modalités de la variable étudiée en notant le nom de chaque modalité sous cette ligne (voir les figures 1 et 2 correspondant respectivement aux exemples 1 et 2 du tri à plat, paragraphe 1.). Les modalités doivent être présentées dans l'ordre croissant (de gauche à droite) dans le cas d'échelles ordinales ou d'intervalles. Notons que, si la variable étudiée est une variable d'intervalles, il est souhaitable de respecter l'échelle lors du marquage des bornes des intervalles.

Puis, on trace un bâton vertical au-dessus de chaque modalité dont la hauteur correspond à leur effectif. La taille des bâtons correspond à l'échelle choisie présentée sur la gauche du graphique par un axe vertical.

On peut éventuellement noter au-dessus de chaque bâton l'effectif correspondant (voir figure 2).

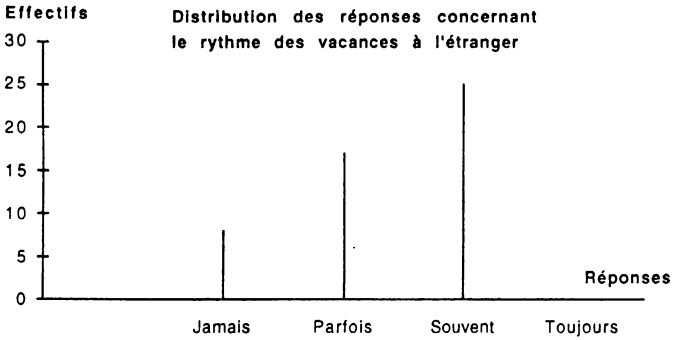


Figure 1

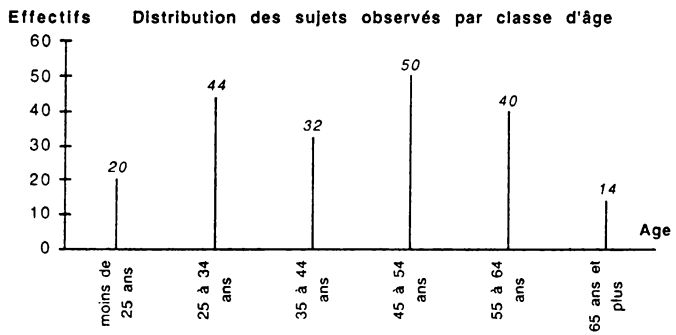


Figure 2

De plus, lors de variables ordinales, on peut représenter l'allure générale de la distribution en reliant les sommets des bâtons (voir figure 3), cela, même en ne représentant pas les bâtons (voir figure 4).

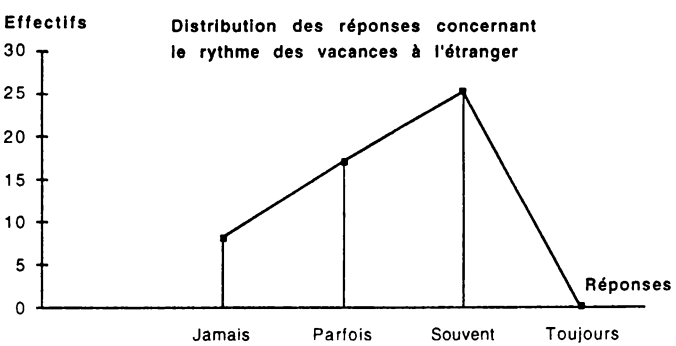


Figure 3

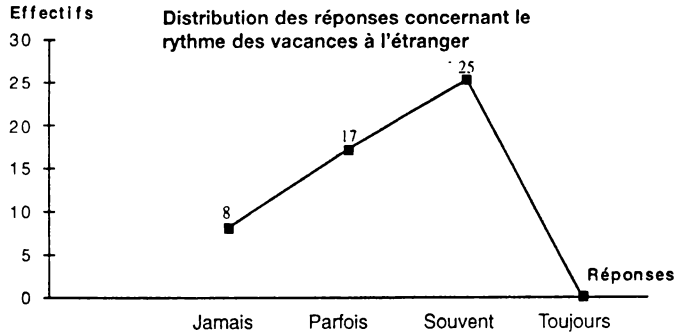


Figure 4

▩ *Remarques :*

Les diagrammes en bâtons peuvent également représenter les distributions des pourcentages ou des fréquences des variables étudiées.

La représentation verticale des diagrammes n'est pas systématique. En effet, dans le cas de la comparaison de deux distributions on peut être amené à une représentation horizontale, tel est le cas pour la pyramide des âges qui permet de comparer le plus souvent les distributions hommes/femmes.

Les diagrammes sont parfois construits à partir de signes plutôt que de bâtons (par exemple des étoiles *) ce qui revient au même pour la lecture de ces graphiques. Dans ce cas, chaque signe correspond à un certain nombre d'observations (par exemple une étoile pour 10 sujets). Ceci est une question d'esthétique et de goût, mais le plus souvent cela dépend de la capacité des logiciels statistiques.

On peut aussi former les bâtons à partir des numéros d'observation de chaque sujet, afin de situer directement les sujets observés individuellement, notamment dans le cas de comparaison de distributions de différentes variables étudiées chez les mêmes individus.

2.2. Histogrammes

2.2.1. Rôle

L'objectif des histogrammes est de visualiser les effectifs de toutes les modalités d'une variable.

2.2.2. Procédure

On utilise la même procédure que pour les diagrammes en bâtons, mais en élargissant les bâtons sur l'intervalle de chaque modalité dans le cas de

variables d'intervalles ou d'une largeur équivalente dans le cas de variables qualitatives (variables nominales). La surface des rectangles ainsi obtenus est proportionnelle aux effectifs de chaque modalité de la variable étudiée (voir les figures 5 et 6 correspondant respectivement aux exemples 1 et 2 du tri à plat, paragraphe 1.).

Notons que l'intitulé de chaque modalité doit figurer de façon centrée en dessous du rectangle correspondant, et non pas sous la marque de l'une des deux bornes des modalités.

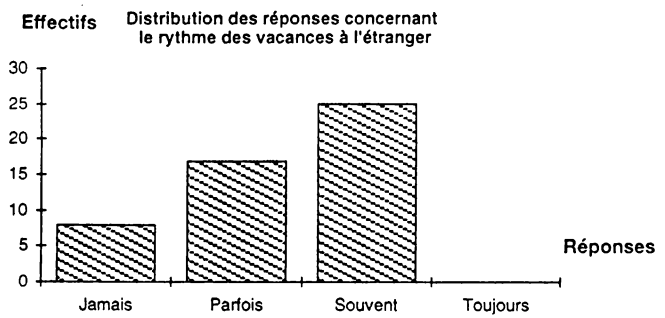


Figure 5

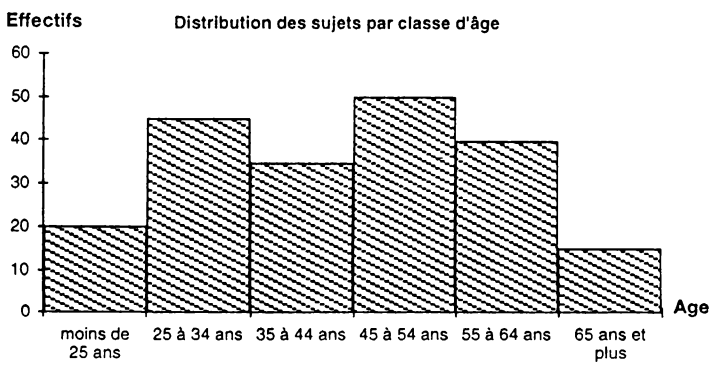


Figure 6

Les histogrammes ont l'avantage de pouvoir représenter clairement les distributions de plusieurs variables ayant les mêmes modalités en les juxtaposant, par exemple dans le cas de plusieurs groupes (voir la figure 7 où l'on distingue la distribution des hommes et des femmes par classe d'âge).

• Exemple :

On a interrogé 200 personnes sur leur âge et on a noté leur réponse par classe d'âge, en spécifiant leur sexe. Les données sont présentées dans le tableau 5 et représentées graphiquement par un histogramme (voir figure 7).

Classes d'âge	moins de 25 ans	25 à 34 ans	35 à 44 ans	45 à 54 ans	55 à 64 ans	65 et plus	Total
Hommes	10	25	15	30	18	5	103
Femmes	10	19	17	20	22	9	97
Total	20	44	32	50	40	14	200

Tableau 5

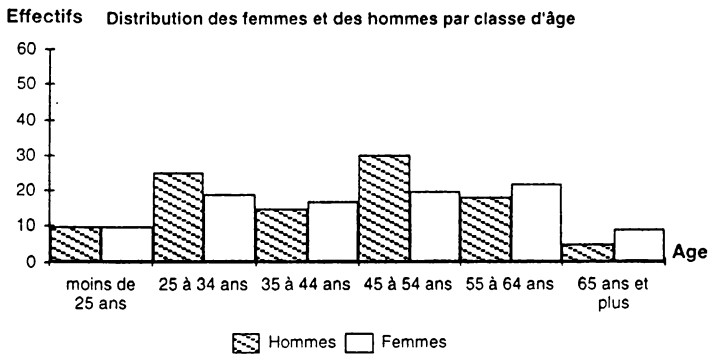


Figure 7

2.3. Camemberts

2.3.1. Rôle

L'objectif des camemberts (ou diagrammes à secteurs) est de **visualiser les effectifs de toutes les modalités d'une variable** ; puis de comparer visuellement plusieurs distributions dans le cas de plusieurs groupes, en juxtaposant les camemberts (voir figure 9).

2.3.2. Procédure

On répartit dans un cercle la distribution de la variable étudiée en découpant ce cercle en autant de parts (ou de quartiers) qu'il y a de modalités, comme lors du découpage d'un camembert (ou d'une tarte). La surface de chaque part doit être proportionnelle à l'effectif de la modalité correspondante. La surface totale du cercle, dit camembert, correspond à l'effectif total (ou à la fréquence 1). Les 360 degrés du cercle correspondent alors à

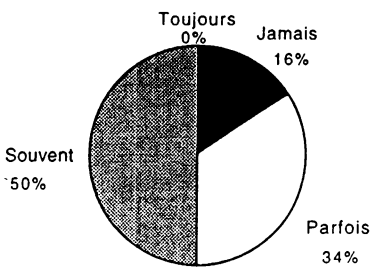
100 % des effectifs, qu'il suffit de répartir selon le pourcentage de chaque modalité (voir les figures 8 et 9 correspondant respectivement aux exemples 1 et 2 du tri à plat, paragraphe 1.). Dans le cas où une modalité a un effectif nul, celle-ci n'apparaît pas sur le camembert.

• Exemple 1 :

Au tri à plat présenté dans le tableau 6 suivant, correspond le camembert présenté figure 8.

Vacances à l'étranger	Jamais	Parfois	Souvent	Toujours	Total
Effectifs	8	17	25	0	50
Fréquences	0,16	0,34	0,5	0	1
Pourcentage	16 %	34 %	50 %	0	100

Tableau 6



Distribution des réponses sur le rythme des vacances prises à l'étranger

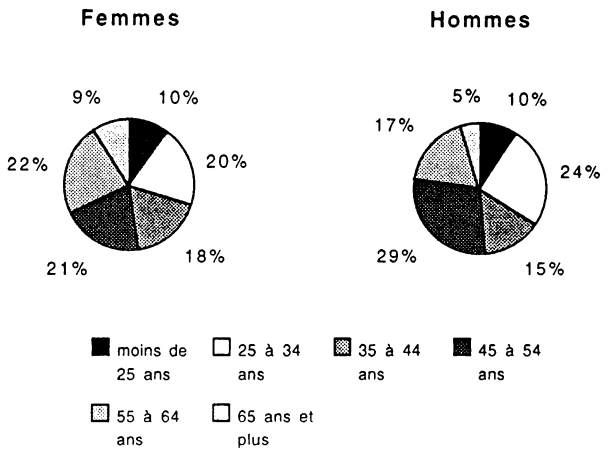
Figure 8

• Exemple 2 :

Aux données représentées dans le tableau 7 suivant, correspondent les camemberts présentés figure 9.

Classes d'âge	moins de 25 ans	25 à 34 ans	35 à 44 ans	45 à 54 ans	55 à 64 ans	65 et plus	Total
Hommes	10	25	15	30	18	5	103
Femmes	10	19	17	20	22	9	97
Total	20	44	32	50	40	14	200

Tableau 7



Distribution des observations par classe d'âge selon le sexe

Figure 9

CHAPITRE V

Statistiques inférentielles

1. Statistique de décision : Tests d'hypothèses *135*
1.1. Test bilatéral *136*
1.2. Test unilatéral à droite *138*
1.3. Test unilatéral à gauche *139*
1.4. Risques d'erreur et puissance d'un test *141*
1.4.1. Risque d'erreur de première espèce : α (alpha) *141*
1.4.2. Risque d'erreur de deuxième espèce : ß (beta) *142*
1.4.3. Puissance du test *142*

2. Tests paramétriques ou tests non paramétriques ? *143*

3. Test de Kolmogorov-Smirnov *144*
3.1. Cas de petits échantillons $5 < n_A < 30$ et $5 < n_B < 30$ *144*
3.1.1. Conditions d'application *144*
3.1.2. Principe du test *145*
3.2. Cas de grands échantillons $n_A \geq 30$ et $n_B \geq 30$ *146*
3.2.1. Conditions d'application *146*
3.2.2. Principe du test *146*

CHAPITRE V

Statistiques inférentielles

L'inférence statistique est la procédure selon laquelle on conclut à des **différences ou non entre des populations** à partir d'observations, c'est-à-dire à partir d'estimations de leurs caractéristiques issues d'échantillons.

1. Statistique de décision : Tests d'hypothèses

Les problèmes que l'on se pose en statistique de décision sont de différents types. Il s'agit :

– soit de ***comparer une population théorique à un échantillon*** afin de déterminer si l'échantillon est issu ou non de cette population de **référence**, c'est-à-dire si la variable estimée à partir de l'échantillon est de même loi statistique ou non que celle de la variable théorique de la population de référence ;

– soit de ***comparer deux ou plus de deux échantillons***, afin de déterminer si ces échantillons sont issus d'une même population ou non ; ou si ces échantillons sont de même loi ou non concernant une variable X, quantitative ou qualitative ;

– soit encore d'***étudier les liens entre les variables étudiées***, qu'elles soient qualitatives ou semi-quantitatives d'une part, ou quantitatives d'autre part.

Afin de répondre à une des questions de ce type, on pose alors une hypothèse, dite ***hypothèse nulle***, que l'on teste contre l'hypothèse contraire, dite ***hypothèse alternative***.

L'hypothèse nulle **correspond à l'hypothèse d'égalité**, c'est-à-dire l'hypothèse selon laquelle par exemple deux moyennes observées sont équivalentes, deux distributions sont équivalentes, ou plusieurs moyennes sont équivalentes. L'hypothèse nulle est notée H_0.

Les statistiques inférentielles permettent de conclure au rejet ou au non rejet de l'hypothèse nulle.

L'hypothèse alternative **choisie** (notée H_1 ou H_a) **correspond à une hypothèse permettant de tester des différences**. On distingue trois types de différences : la **différence simple**, l'**inégalité à droite** et l'**inégalité à gauche**, qui correspondent respectivement aux tests bilatéral, unilatéral à droite et unilatéral à gauche. **Le test bilatéral est le test le plus utilisé** car la question principale que l'on se pose en statistiques inférentielles est de savoir si les échantillons observés sont issus ou non d'une même population de référence, ou encore s'il existe une liaison ou non entre deux variables.

1.1. Test bilatéral

Lors d'un *test bilatéral* on cherche à savoir s'il existe une différence entre les paramètres comparés. Alors, le risque α de rejeter l'hypothèse nulle à tort se situe aux extrêmes de part et d'autre de la distribution sous H_0 (voir les figures 1, 2, 3 et 4 dans le cas de lois normales).

Par exemple, quand on teste la différence entre deux moyennes observées m_1 et m_2, on se demande ***si les moyennes vraies des deux populations de référence***, dont sont issus les échantillons observés, ***sont équivalentes ou non***.

Les hypothèses à tester sont alors : $\begin{cases} H_0 : \mu_1 = \mu_2 \\ H_1 : \mu_1 \neq \mu_2 \end{cases}$

Ce qui revient à tester : $\begin{cases} H_0 : \mu_1 - \mu_2 = 0 \\ H_1 : \mu_1 - \mu_2 \neq 0 \end{cases}$

Peu importe de savoir si l'hypothèse alternative se situe à droite ou à gauche de l'hypothèse nulle, on recherche simplement l'existence ou non d'une différence entre les moyennes théoriques à partir des moyennes observées.

L'hypothèse nulle suppose dans cet exemple que les moyennes μ_1 et μ_2 sont équivalentes à une moyenne théorique commune μ.

Soit : $\begin{cases} H_0 : \mu_1 = \mu_2 = \mu \\ H_1 : \mu_1 \neq \mu_2 \end{cases}$

Dans le cas d'une différence significative, on a par exemple :

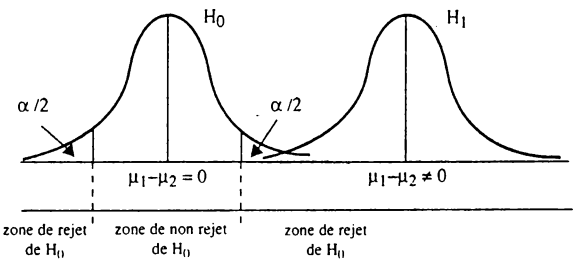


Figure 1

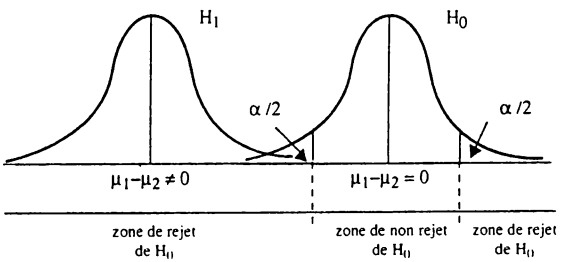


Figure 2

Dans le cas d'une différence non significative, on a par exemple :

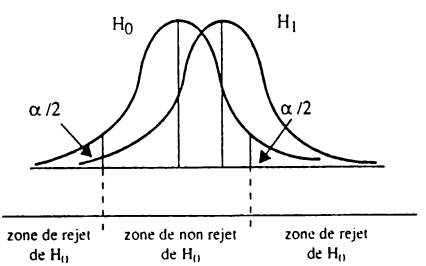


Figure 3

soit

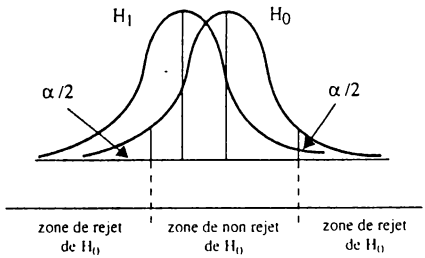


Figure 4

Dans le cas du test du χ^2, seul un test bilatéral est envisagé. La question est de savoir si les distributions diffèrent ou non.

1.2. Test unilatéral à droite

Dans le cas du *test unilatéral à droite*, on se préoccupe de savoir s'il existe une inégalité à droite entre les paramètres comparés concernant une variable X. Alors, le risque α de rejeter l'hypothèse nulle à tort se situe vers la droite de la distribution de H_0 (voir les figures 5 et 6). Par exemple, quand on teste l'inégalité entre deux moyennes observées m_1 et m_2 (à partir de 2 échantillons), on se demande ***si la moyenne vraie de la première population de référence est supérieure ou non à celle de la deuxième population*** (les échantillons étant issus des populations de référence).

Les hypothèses à tester sont alors :

$$\begin{cases} H_0 : \mu_1 = \mu_2 = \mu \\ H_1 : \mu_1 > \mu_2 \end{cases}$$

Ce qui revient à tester :

$$\begin{cases} H_0 : \mu_1 - \mu_2 = 0 \\ H_1 : \mu_1 - \mu_2 > 0 \end{cases}$$

■ **Un moyen mnémotechnique** pour se rappeler du sens de l'inégalité selon le test est d'associer la forme du signe > à D comme Droite, puis d'associer la forme du signe < à G comme Gauche.

Il importe alors de savoir si l'hypothèse alternative se situe à droite de l'hypothèse nulle ou non.

Dans le cas d'une inégalité à droite significative, on a par exemple :

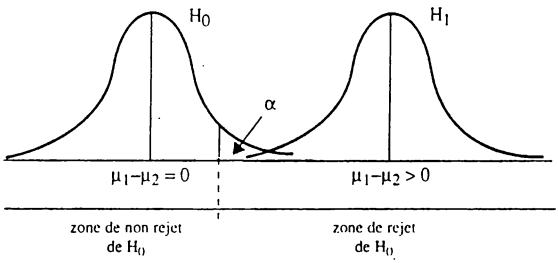


Figure 5

Dans le cas d'une différence non significative, on a par exemple :

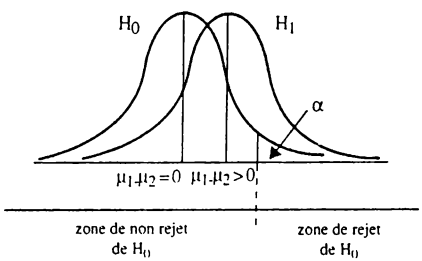


Figure 6

1.3. Test unilatéral à gauche

Dans le cas du ***test unilatéral à gauche*** on se préoccupe de savoir s'il existe une inégalité à gauche entre les paramètres comparés. Alors, le risque α de rejeter l'hypothèse nulle se situe à gauche de la distribution de H_0 (voir les figures 7 et 8). Par exemple, quand on teste l'inégalité entre deux moyennes observées m_1 et m_2, on se demande ***si la moyenne vraie de la première population de référence μ_1 est inférieure ou non à celle de la deuxième population μ_2*** (les échantillons étant issus des populations de référence).

Les hypothèses sont alors : $\begin{cases} H_0 : \mu_1 = \mu_2 \\ H_1 : \mu_1 < \mu_2 \end{cases}$

Il importe alors de savoir si l'hypothèse alternative se situe à gauche de l'hypothèse nulle ou non.

Dans le cas d'une inégalité à gauche significative, on a par exemple :

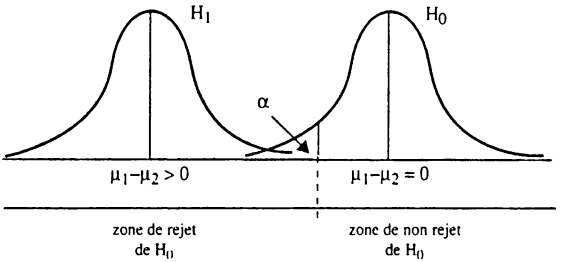


Figure 7

Dans le cas d'une différence non significative, on a par exemple :

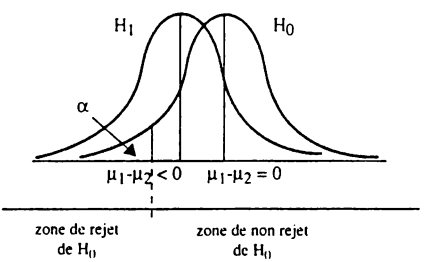
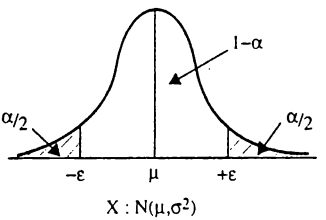


Figure 8

1.4. Risques d'erreur et puissance d'un test

1.4.1. Risque d'erreur de première espèce : α (alpha)

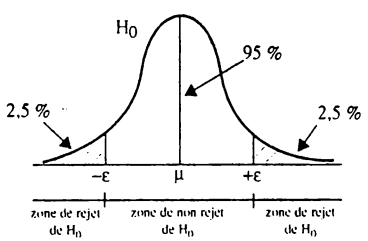
Dans le cas où la variable étudiée X d'une population donnée est distribuée normalement, de moyenne μ et de variance σ^2 (par exemple la taille des individus) on pourra représenter graphiquement la distribution de cette variable ainsi :



X : N(μ,σ^2)

Figure 9

Dans la plupart des cas, la majorité des individus se situe autour de la moyenne μ, mais certains s'en éloignent. Or plus les individus s'éloignent de la moyenne, plus ils diffèrent de la majorité des individus (selon le caractère étudié). A partir d'un certain seuil, on dira qu'ils diffèrent significativement de l'ensemble des sujets. Le seuil fixé par consensus par les statisticiens est de 5 %. C'est-à-dire que l'on considère que 5 % des individus les plus éloignés de la moyenne sont "hors normes", ou encore ne sont pas représentatifs de la distribution ou de l'ensemble des sujets. Ce seuil peut aussi être fixé arbitrairement, par exemple à 1 % ou 10 %. Il correspond au risque d'erreur de première espèce, soit α (correspondant à la surface hachurée de la figure 9 ci-dessus). **C'est le risque de se tromper en rejetant l'hypothèse nulle**. Ainsi, le risque α est toujours défini à partir de l'hypothèse nulle H_0. C'est pourquoi ***α correspond à la zone de rejet de l'hypothèse nulle sous la courbe H_0*** et peut être représenté ainsi (figure 10) quand il est égal à 5 %, dans le cas d'un test bilatéral :



X : N(μ,σ^2)

Figure 10

Le risque d'erreur α correspond à la probabilité de se tromper quand on rejette H_0, c'est-à-dire à la ***probabilité de rejeter H_0, alors que H_0 est vraie*** (c'est-à-dire rejet de H_0 à tort).

Soit α = risque d'erreur de première espèce
= P(X ∈ zone de rejet de H_0 / H_0 vraie)
= P(rejet de H_0 / H_0 vraie)
= P(non rejet de H_1 / sous l'hypothèse H_0)

1.4.2. Risque d'erreur de deuxième espèce : ß (beta)

Un second risque d'erreur est envisageable, celui de se tromper quand on rejette H_1, c'est-à-dire ***la probabilité de rejeter H_1, alors que H_1 est vraie*** (c'est-à-dire rejet de H_1 à tort).

On parle alors du risque d'erreur de deuxième espèce, que l'on note ß, et qui se définit sous l'hypothèse H_1 (voir figure 11).

La figure 11 représente le test de l'hypothèse H_0 contre l'hypothèse H_1, soit : $\begin{cases} H_0 : \mu_1 = \mu_2 \\ H_1 : \mu_1 \neq \mu_2 \end{cases}$

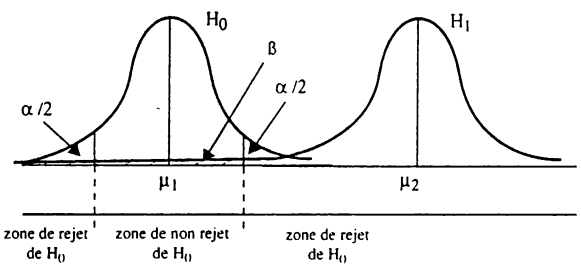


Figure 11

Alors ß = risque de deuxième espèce (rejet de H_1 à tort)
= P(X ∈ zone de rejet de H_1 / H_1 vraie)
= P(rejet de H_1 / H_1 vraie)
= P(X ∈ zone de non rejet de H_0 / sous l'hypothèse H_1 vraie)
= P(non rejet de H_0 / H_1 vraie)

1.4.3. Puissance du test

La probabilité complémentaire du risque ß correspond à la puissance du test, notée P, c'est-à-dire à la ***probabilité de non rejet de l'hypothèse H_1 à juste raison*** (voir figure 12).

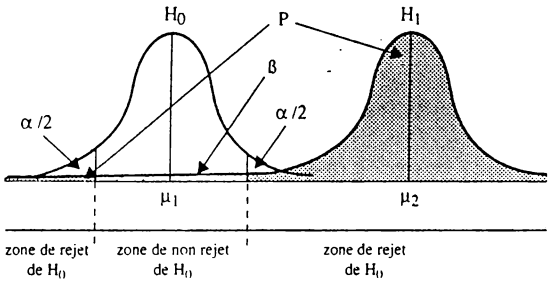


Figure 12

Soit P = Puissance du test
= 1 - β
= 1 - P(rejet de H_1 / H_1 vraie)
= P(non rejet de H_1 / H_1 vraie)
= P(X ∈ zone de rejet de H_0 / sous l'hypothèse H_1 vraie)

▩ *Remarque :*
Plus la taille de l'échantillon est grande, plus la puissance du test augmente.

2. Tests paramétriques ou tests non paramétriques ?

Les tests d'hypothèses basés sur des données observées (issues d'un ou plusieurs échantillons décrivant une ou plusieurs variables) permettent de déduire les paramètres inconnus de la population de référence.

Or, lors des **tests paramétriques**, ces inférences ne sont possibles que selon **certaines conditions d'applications**.

On distingue les tests paramétriques des tests non paramétriques par le simple fait que les tests paramétriques posent des hypothèses sur les distributions observées, alors qu'aucune contrainte n'est imposée aux distributions observées lors des tests non paramétriques (hormis les effectifs minimum nécessaires).

C'est pourquoi le test du Khi-deux (χ^2) est considéré comme un test non paramétrique (aucune contrainte n'est imposée aux distributions observées mais effectifs supérieurs ou égaux à 5).

La condition d'application la plus fréquente exige que la ou les varia-

bles étudiées soient distribuées normalement, c'est-à-dire suivent une loi normale. **Un des moyens de déterminer si une distribution est normale** (conforme au modèle de normalité de Laplace-Gauss) **est de faire un test de Kolmogorov-Smirnov**. Mais, quand cette condition d'application n'est pas vérifiée, ou quand les effectifs sont trop faibles, alors des **tests non paramétriques, moins puissants**, permettent néanmoins de tester les hypothèses dans le but d'une inférence. L'arbre de décision présenté dans le chapitre I distingue les tests paramétriques des tests non paramétriques.

3. Test de Kolmogorov-Smirnov

Le test de Kolmogorov-Smirnov permet de comparer deux distributions. Afin de **tester la normalité d'une distribution** on peut être amené à faire ce test.

▩ ***Rôle du test :***

Comparaison de deux distributions continues de 2 caractères X et X' de 2 échantillons A et B **indépendants**, dans le cas de *séries classées*. L'une des deux distributions peut être normale *a priori* afin de tester la normalité de l'autre distribution.

La procédure du test est indiquée ici à partir d'exemples, selon le cas de petits ou grands échantillons.

Le test du χ^2 peut aussi être réalisé pour comparer deux distributions, dans le cas de séries classées, *si tous les effectifs sont supérieurs ou égaux à 5*.

3.1. Cas de petits échantillons $5 < n_A < 30$ et $5 < n_B < 30$

3.1.1. Conditions d'application

1) variable continue et distribution quelconque dans chaque population.
2) mêmes limites de classes pour chacun des échantillons, ces limites pouvant être arbitraires
3) éviter $n < 5$
4) $n_A = n_B$

Hypothèses :

H_0 : (A) = (B), les lois des échantillons A et B sont équivalentes

contre
- H_1 : (A) ≠ (B) Test bilatéral, les lois sont différentes
- ou H_1': (A) > (B) Test unilatéral à droite
- ou H_1'': (A) < (B) Test unilatéral à gauche

3.1.2. Principe du test

On calcule les effectifs cumulés S_A et S_B, et leur différence $S_A - S_B$, en gardant en mémoire le signe de cette différence pour la suite du test.

Soit $$D_c = \max |S_A - S_B|$$

avec D_c la valeur maximale des différences $S_A - S_B$ en valeur absolue. On la note D_c (D calculé) pour la différencier de D théorique lu dans la table (dit aussi D tabulé) noté D_T.

Puis on lit la valeur de D_T dans la table de Kolmogorov au seuil $\alpha = 5$ % (voir table en annexe dans le cas de petits échantillons).

Si $D_c \geq D_T$ alors on rejette l'hypothèse H_0.

Dans le cas où l'on rejette H_0, on recherche le degré ou seuil de signification de la différence observée dans la table. C'est-à-dire que l'on se pose la question suivante : "Jusqu'à quelle valeur de α puis-je encore conclure à une différence significative, cette probabilité étant notée p ?" Si cela correspond à $\alpha = 1$ %, on note $p < 0{,}01$.

• Exemple :

On souhaite déterminer si les distributions des échantillons A et B sont équivalentes ou non.

On note alors dans un tableau les effectifs observés dans chaque classe de chacun des échantillons, puis les effectifs cumulés de chacun des échantillons, et la différence entre les deux échantillons des effectifs cumulés pour chaque classe. Soit :

limites de classe	14	16	18	20	22	24	26	28	30
éch. A		0	1	1	2	5	3	2	1
éch. B		0	0	1	4	6	2	2	0
S_A		-	1	2	4	9	12	14	15
S_B		-	0	1	5	11	13	15	15
$S_A - S_B$			1	1	-1	-2	-1	-1	0

Tableau 1

Avec S_A et S_B les effectifs cumulés des échantillons A et B.

alors $D_c = \max |S_A - S_B| = |D_c| = 2$

si $D_c \geq D_T$ alors on rejette l'hypothèse H_0. Or $D_c = 2$ est inférieur à $D_T = 8$ au seuil $\alpha = 5$ %, donc on ne rejette pas H_0, c'est-à-dire que la différence est non significative (NS).

3.2. Cas de grands échantillons $n_A \geq 30$ et $n_B \geq 30$

3.2.1. Conditions d'application

1) variable continue et distribution quelconque dans chaque population.
2) mêmes limites de classes pour chacun des échantillons, ces limites pouvant être arbitraires
3) $n_A \geq 30$ et $n_B \geq 30$ avec $n_A = n_B$ ou $n_A \neq n_B$

Hypothèses :

H_0 : $(A) = (B)$

contre
- H_1 : $(A) \neq (B)$ Test bilatéral
- ou H_1' : $(A) > (B)$ Test unilatéral à droite
- ou H_1'' : $(A) < (B)$ Test unilatéral à gauche

3.2.2. Principe du test

On calcule les effectifs cumulés S_A et S_B et les fréquences cumulées F_A et F_B. Le critère est la différence entre ces fréquences : $F_A - F_B$, et

$$d = \max |F_A - F_B|$$

en conservant le signe où d est la valeur maximale des différences des fréquences cumulées $F_A - F_B$ en valeur absolue.

On calcule :

$$c_c = d\sqrt{\frac{n_A n_B}{n_A + n_B}}$$

et on compare la valeur c_c (paramètre calculé) à c_T (paramètre tabulé) lu dans la table de Kolmogorov, en choisissant la table dans le cas de grands échantillons et selon le test bilatéral ou unilatéral (voir annexe).

Dans le cas d'un test bilatéral, au seuil $\alpha = 5$ % on a toujours $c_T = 1{,}36$. Si $|c_c| > c_T$ on rejette l'hypothèse H_0.

Dans le cas d'un test unilatéral, au seuil $\alpha = 5$ % on a toujours $c_T = 1{,}22$.

Pour un test unilatéral à gauche (H_1'), si $c_c < 0$ et si $c_c < -c_T$ on rejette l'hypothèse H_0.

Pour un test unilatéral à droite (H_1''), si $c_c < 0$ on rejette l'hypothèse H_0.

Dans le cas où on rejette H_0, on recherche le degré ou seuil de signification de la différence observée dans la table. C'est-à-dire que l'on se pose la question suivante : jusqu'à quelle valeur de α puis-je encore conclure à une différence significative, cette probabilité étant notée p ? Si cela correspond à $\alpha = 1$ %, on note $p < 0{,}01$.

• Exemple :

On souhaite déterminer si les distributions des échantillons A et B sont équivalentes ou non.

On note dans un tableau les effectifs observés dans chaque classe de chacun des échantillons (éch. A et éch. B), puis les effectifs cumulés de chacun des échantillons (S_A et S_B), puis les fréquences cumulées de chacun des échantillons (F_A et F_B) et la différence entre les deux échantillons des fréquences cumulées pour chaque classe (F_A - F_B), soit :

lim cl.	14	16	18	20	22	24	26	28	30	32	34	36
éch. A	2	1	11	14	23	10	19	7	6	6	1	
éch. B	0	0	4	10	34	20	2	0	0	0	0	
S_A	2	3	14	28	51	61	80	87	93	99	100	
S_B	0	0	4	14	48	68	70	70	70	70	70	
F_A %	2	3	14	28	51	61	80	87	93	99	100	
F_B %	0	0	6	20	69	97	100	100	100	100	100	
F_A - F_B	+2	+3	+8	+8	-18	-36	-20	-13	- 7	- 1	0	

Tableau 2

Alors $d = \max F_A - F_B = -36$ (en conservant le signe)

$$c_c = d\sqrt{\frac{n_A n_B}{n_A + n_B}} = -36\sqrt{\frac{100 \times 70}{170}} = -2{,}31$$

La valeur de c_T est lue dans la table de Kolmogorov-Smirnov, dans le cas de grands échantillons.

Test bilatéral :

α	c_T	
5%	1,36	$\Rightarrow \lvert c_c \rvert = 2{,}31 > c_T \Rightarrow$ on rejette $H_0 \Rightarrow$ différence Significative (S)
1%	1,63	

Donc on peut dire que les deux distributions sont différentes significativement avec un degré de signification $p < 0{,}001$, car pour $\alpha = 1$ ‰, on lit $c_T = 1{,}95$ (dans le cas d'un test bilatéral).

Test unilatéral : il faut tenir compte du signe de D

H_1' : (A) < (B) $\Rightarrow$ -2,31 < -1,22 $\Rightarrow$ (A) < (B) $\Rightarrow$ on rejette l'hypothèse $H_0 \Rightarrow$ S

H_1'' : (A) > (B) $\Rightarrow c_c < 0 \Rightarrow$ on rejette l'hypothèse $H_1'' \Rightarrow$ NS

$\Rightarrow$ NS (inégalité Non Significative)

CHAPITRE VI

Tests des différences entre variables qualitatives à deux ou n classes

1. Données non appariées *151*
1.1. Test du χ^2 ..... *151*
1.1.1. Comparaison d'une répartition observée à une répartition théorique *152*
1.1.2. Comparaison de plusieurs répartitions observées *156*
1.2. Le χ^2 corrigé de Yates (tableau 2x2) *159*
1.2.1. Rôle du test *159*
1.2.2. Principe du test *160*

2. Données appariées - χ^2 de Mac Nemar *162*

3. Récapitulatif *164*

4. Exercices *166*
4.1. Exercices corrigés *166*
4.2. Exercices non corrigés *167*
4.3. Solutions *169*

CHAPITRE VI

Tests des différences entre variables qualitatives à deux ou n classes

1. Données non appariées

1.1. Test du χ^2

▩ *Rôle du test :*

Le test du χ^2 (χ est la lettre grecque Khi), ou du Khi-carré (noté aussi parfois chi-carré), permet de comparer les répartitions d'échantillons indépendants concernant une variable **qualitative** (ou **semi-quantitative**, c'est-à-dire une variable à plusieurs classes), et la liaison entre deux variables qualitatives dans le cas d'un seul échantillon cross-classifié.

Autrement dit, le test du χ^2 permet de comparer :

– **soit une répartition observée à une répartition théorique**, par exemple la répartition homme/femme parmi les députés à l'Assemblée nationale en France en 1998 comparativement à la théorie d'égalité des chances correspondant à 50 % de femmes et 50 % d'hommes.

– **soit plusieurs répartitions observées**, c'est-à-dire les distributions de plusieurs échantillons ; par exemple la répartition homme/femme parmi les députés à l'Assemblée nationale en France en 1998, comparativement à celle de 1996.

– **soit plusieurs répartitions observées issues d'un seul échantillon**, selon les différentes modalités d'une variable qualitative. Dans ce cas, le test du χ^2 permet d'étudier la liaison ou l'indépendance entre deux caractères qualitatifs. Par exemple, concernant l'échantillon des sujets inscrits à un parti politique, on peut souhaiter étudier la liaison ou l'indépendance entre la variable "appartenance à un parti politique donné" (variable

qualitative à deux classes si l'on considère les partis de droite d'une part, et les partis de gauche d'autre part) et la variable "sexe" (variable qualitative à deux classes : féminin, masculin) afin de déterminer si le choix d'un mouvement politique dépend ou non du sexe, pour une population donnée.

Le χ^2 mesure la distance entre les effectifs réellement observés et les effectifs théoriques (ou calculés, c'est-à-dire les effectifs qui seraient observés si H_0 était vraie). Le test du χ^2 ne posant pas d'hypothèses sur la forme des distributions étudiées, il est considéré comme faisant partie des **tests non paramétriques.**

1.1.1. Comparaison d'une répartition observée à une répartition théorique

■ ***Procédure :***

On étudie un caractère qualitatif X à k classes, dont les répartitions observées et théoriques des effectifs par classe peuvent être présentées dans le tableau 1 suivant :

	classe 1	classe 2	classe 3	...	classe k	Total
Effectifs observés (O_i)	O_1	O_2	O_3		O_k	$\sum_{i=1}^{k} O_i$
Proportions théoriques (E_i)	P_1	P_2	P_3		P_k	1
Effectifs théoriques (C_i)	C_1	C_2	C_3		C_k	$\sum_{i=1}^{k} C_i$

Tableau 1

Afin de rendre ces deux répartitions comparables, les effectifs théoriques doivent être calculés de façon à ce que le total des effectifs théoriques soit équivalent au total des effectifs observés (sans jamais modifier les effectifs observés), soit :

$$\sum_{i=1}^{k} C_i = \sum_{i=1}^{k} O_i$$

Les hypothèses à tester sont :

H_0 : La répartition observée du caractère X est équivalente à la répartition théorique de ce caractère X

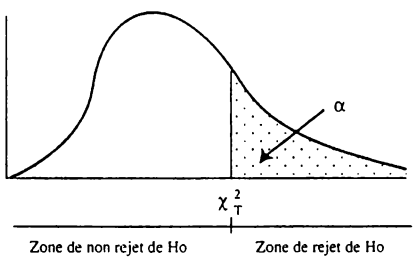
H_1 : La répartition observée du caractère X est différente de la répartition théorique de ce caractère X

Pour comparer une répartition observée à une répartition théorique à k classes, on utilise le test du χ^2 et on détermine le paramètre calculé χ^2_c :

$$\chi^2_c = \sum_{i=1}^{k} \frac{(O_i - C_i)^2}{C_i}$$

avec O_i = effectifs observés,
C_i = effectifs calculés.

Puis, on cherche le χ^2 tabulé, noté χ^2_T, lu dans la table du χ^2 (jointe en annexe), pour le risque d'erreur de première espèce α (le plus souvent $\alpha = 5\ \%$) et pour un degré de liberté égal à k-1 (c'est-à-dire le nombre de classes moins un). Par exemple, pour $\alpha = 5\ \% = 0{,}05$ et pour un degré de liberté égal à 1 (ddl = 1), la valeur du χ^2 tabulé est $\chi^2_T = 3{,}84$.



Si $\chi^2_c < \chi^2_T$ alors on rejette l'hypothèse alternative H_1 et on conclut à une différence non significative entre la répartition observée et la répartition théorique.

Si $\chi^2_c \geq \chi^2_T$ alors on rejette l'hypothèse nulle et on conclut à une différence significative entre la répartition observée et la répartition théorique. La valeur de α lue dans la table correspondant au χ^2_T le plus proche et inférieur au χ^2_c fixe le degré de signification.

Conditions d'application :

Tous les effectifs calculés doivent être supérieurs ou égaux à 5 (tous les $C_i \geq 5$).

• Exemple :

On dispose d'une population de 2000 sujets répartis en quatre classes d'âge dont la composition est présentée dans le tableau 2 suivant. Un échantillon de 200 sujets a été constitué à partir de cette population (dite de référence). Afin de déterminer si cet échantillon est représentatif de la population initiale, on compare la répartition observée obtenue à celle de la population théorique de référence, à l'aide d'un test du χ^2.

	moins de 20 ans	20 à 39 ans	40 à 59 ans	60 ans et plus	Total
Echantillon observé (O_i)	70	40	68	22	$\sum_{i=1}^{k} O_i = 200$
Population générale (E_i)	600	400	800	200	$\sum_{i=1}^{k} E_i = 2000$
Répartition théorique (C_i)	60	40	80	20	$\sum_{i=1}^{k} C_i = 200$

Tableau 2

On veut vérifier si la répartition observée est conforme à la répartition théorique, les hypothèses sont alors :

H_0 : la répartition observée des sujets selon la classe d'âge est équivalente à la répartition théorique des sujets (selon la classe d'âge)

H_1 : la répartition observée est différente de la répartition théorique.

Les conditions d'application sont retrouvées : Les effectifs calculés sont tous supérieurs à 5.

On calcule la valeur du paramètre du χ^2 :

$$\chi^2_c = \sum_{i=1}^{k} \frac{(O_i - C_i)^2}{C_i}$$

Dans le tableau 2, on ramène l'effectif théorique à 200 (car 200 sujets sont observés dans l'échantillon). On note les effectifs calculés C_i et les effectifs observés O_i :

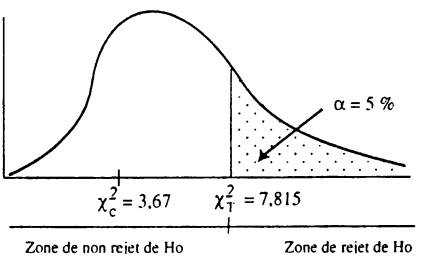
$$\chi^2_c = \frac{(O_1 - C_1)^2}{C_1} + \frac{(O_2 - C_2)^2}{C_2} + \frac{(O_3 - C_3)^2}{C_3} + \frac{(O_4 - C_4)^2}{C_4}$$

$$\chi^2_c = \frac{10^2}{60} + \frac{0^2}{40} + \frac{(-12)^2}{80} + \frac{2^2}{20}$$

$$\chi^2_c = \frac{100}{60} + \frac{144}{80} + \frac{4}{20}$$

$$\chi^2_c = 3{,}67$$

On compare la valeur du χ^2_C à celle du χ^2_T lue dans la table du χ^2 à k-1 degrés de liberté, c'est-à-dire avec un ddl = 3.
Dans la table du χ^2, on lit dans la colonne $\alpha_T = 0{,}05$ la valeur correspondante du χ^2 tabulé à 3 ddl, soit $\chi^2_T = 7{,}815$.



Ainsi, on a $\chi^2_C < \chi^2_T$, c'est-à-dire que le χ^2 calculé se situe dans la zone de non rejet de l'hypothèse nulle, donc la différence entre les répartitions observée et théorique est non significative.
On conclut alors que la répartition observée des sujets selon la classe d'âge est équivalente à la répartition théorique des sujets selon la classe d'âge (dans la population générale de référence). L'échantillon observé est alors représentatif de la population de référence concernant sa répartition par classe d'âge, au risque $\alpha = 5$ %.

1.1.2. Comparaison de plusieurs répartitions observées

Lors de la comparaison de plusieurs répartitions observées, le test du χ^2 permet :

– **de déterminer si les répartitions observées sont équivalentes** (c'est-à-dire si les échantillons sont issus d'une même population de référence) ;

– ou **d'étudier la liaison entre les deux caractères qualitatifs dans le cas où l'on ne dispose que d'un seul échantillon, mais classé en plusieurs groupes selon une variable donnée**. Les deux variables étudiées figurent dans le tableau de contingence des données observées (autrement dit, les deux entrées du tableau correspondent aux deux caractères qualitatifs dont on étudie la liaison).

Quand on extrait plusieurs échantillons d'une population de référence à C classes, on admet que la somme des χ^2 de chaque échantillon suit la loi du χ^2 pour la somme des degrés de liberté.

• Procédure à partir d'un exemple :

On compare la production de publications de trois unités de recherche d'un département de psychologie d'une université, sur les quatre dernières années. Les données observées sont présentées dans le tableau de contingence suivant :

Type de publications / Type d'unité	Articles dans des revues à comité de lecture	Articles dans des revues sans comité de lecture	Communications référencées dans des congrès	Total
psychologie sociale	80	10	40	130
psychologie du développement	60	10	20	90
psychologie de la santé	120	20	60	200
Total	260	40	120	420

Tableau 3

On compare les répartitions observées des publications des trois unités de recherche étudiées, ce qui revient également à étudier la liaison entre les deux caractères qualitatifs : "discipline de l'unité de recherche" et "type de publications", afin de déterminer si le type de publications est lié ou non à la discipline de recherche.

Les hypothèses à tester sont alors :

$\begin{cases} H_0 : \text{Il n'y a pas de différence entre les répartitions observées de type de publication selon les unités de recherche} \\ H_1 : \text{Il y a une différence entre les répartitions observées} \end{cases}$

ou encore

$\begin{cases} H_0 : \text{Il n'y a pas de liaison entre les caractères observés} \\ H_1 : \text{Il y a une liaison entre les caractères qualitatifs étudiés (soit : les variables sont liées, c'est-à-dire qu'il y a dépendance entre les deux caractères).} \end{cases}$

Pour étudier l'indépendance entre deux caractères qualitatifs à partir d'un tableau de contingence à L lignes et C colonnes, on détermine pour chaque cellule du tableau l'effectif calculé (ou effectif théorique) sous l'hypothèse d'indépendance. Cet effectif théorique est, pour chaque cellule du tableau, le produit des marges divisé par le total général (c'est-à-dire le produit du total de la ligne et du total de la colonne correspondant à la case, divisé par le total général).

Par exemple, pour la première cellule du tableau, c'est-à-dire la case correspondant à la ligne 1 et à la colonne 1, l'effectif observé est 80 (soit 80 publications dans des revues à comité de lecture pour le laboratoire de psychologie sociale), noté $O_{11} = 80$; et l'effectif calculé (ou théorique) correspond au produit des marges divisé par le total général, soit $\frac{130 \times 260}{420} = 80{,}48$; noté $C_{11} = 80{,}48$.

Ce calcul devant être effectué pour chaque cellule du tableau afin de comparer les effectifs observés aux effectifs calculés sous l'hypothèse d'indépendance, il est recommandé d'inscrire les effectifs calculés dans chaque cellule correspondante pour plus de clarté. Dans le tableau 4 suivant, les effectifs calculés sont notés entre parenthèses dans chaque cellule :

Type de publications / Type d'unité	Articles dans des revues à comité de lecture	Articles dans des revues sans comité de lecture	Communications référencées dans des congrès	Total
psychologie sociale	80 (80,48)	10 (12,38)	40 (37,14)	130
psychologie du développement	60 (55,71)	10 (8,57)	20 (25,71)	90
psychologie de la santé	120 (123,81)	20 (19,05)	60 (57,14)	200
Total	260	40	120	420

Tableau 4

■ ***Conditions d'application :***

Tous les effectifs calculés doivent être supérieurs ou égaux à 5 (tous les $C_j \geq 5$). Ces conditions sont respectées dans cet exemple.

Puis, on calcule le χ^2 dont la formule est :

$$\chi^2_c = \sum_{i,j} \frac{(O_{ij} - C_{ij})^2}{C_{ij}}$$

avec $i = 1, 2, ..., L$ (numéro de ligne)
$j = 1, 2, ..., C$ (numéro de colonne)

Soit $$\chi^2_c = \sum_{i=1}^{3} \sum_{j=1}^{3} \frac{(O_{ij} - C_{ij})^2}{C_{ij}}$$

avec $i = 1, 2, 3$
$j = 1, 2, 3$

$$\chi^2_c = \frac{(80-80,48)^2}{80,48} + \frac{(10-12,38)^2}{12,38} + \frac{(40-37,14)^2}{37,14} + \frac{(60-55,71)^2}{55,71}$$
$$+ \frac{(10-8,57)^2}{8,57} + \frac{(20-25,71)^2}{25,71} + \frac{(120-123,81)^2}{123,81} + \frac{(20-19,05)^2}{19,05}$$
$$+ \frac{(60-57,14)^2}{57,14}$$

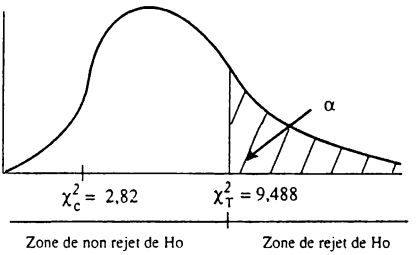
$$\chi^2_c = 2,82$$

On cherche alors, pour le risque α (le plus souvent $\alpha = 5\ \%$), la valeur du χ^2 tabulé dans la table du χ^2, pour un nombre de degrés de liberté égal au nombre de lignes moins un multiplié par le nombre de colonnes moins un.

Soit $ddl = (L - 1) \times (C - 1)$

Dans l'exemple $ddl = (3 - 1) \times (3 - 1) = 4$

Alors pour $\alpha = 5\ \%$ et un ddl = 4, la valeur du χ^2 tabulé est $\chi^2_T = 9,488$.



Si $\chi^2_c < \chi^2_T$, alors on rejette l'hypothèse alternative H_1 et on conclut à une différence non significative entre les répartitions observées, c'est-à-dire à une indépendance entre les deux caractères qualitatifs étudiés.
Si $\chi^2_c \geq \chi^2_T$, alors on rejette l'hypothèse nulle et on conclut à une différence significative entre les répartitions observées, c'est-à-dire à une liaison entre les deux caractères qualitatifs étudiés. Puis, la valeur de α (lue dans la table du χ^2) correspondant au χ^2_T le plus proche et inférieur au χ^2_C, fixe le degré de signification de la différence.
Autrement dit, si le "α" correspondant au χ^2 calculé est supérieur à 5 %, la liaison est significative ; ou si au contraire il est inférieur ou égal à 5 %, la liaison n'est pas significative.
Dans l'exemple étudié on a $\chi^2_C = 2,82$ et $\chi^2_T = 9,488$; donc $\chi^2_C < \chi^2_T$, alors on rejette l'hypothèse alternative H_1 et on conclut à une différence non significative entre les répartitions observées de type de publication selon les laboratoires de psychologie. Ainsi il n'y a pas de liaison entre les deux caractères qualitatifs étudiés, c'est-à-dire que le type de publications produites est indépendant de la discipline des unités de recherche étudiées.

1.2. Le χ^2 corrigé de Yates (tableau 2x2)

1.2.1. Rôle du test

Le χ^2 corrigé de Yates est utilisé pour la comparaison d'effectifs calculés (ou théoriques) et observés (lors de l'utilisation du test du χ^2), **quand un des effectifs calculés est proche de 5, mais supérieur à 2.** D'une façon générale le χ^2 corrigé de Yates est utilisé quand le nombre de degrés de liberté est égal à 1 (ddl = 1), c'est-à-dire lorsque l'on étudie des caractères à 2 classes, donc des tableaux de contingence à 2 lignes et 2 colonnes (tableaux 2x2).

1.2.2. Principe du test

Le principe de cette méthode est de corriger le χ^2 en diminuant la valeur absolue de chaque écart (O - C) de $^1/_2$ afin de réduire la valeur du χ^2 et de ne pas conclure à tort à une différence significative entre les répartitions.

– Dans le cas de la comparaison d'une répartition observée à une répartition théorique à 2 classes,

	Classe 1	Classe 2
Répartition observée	O_1	O_2
Répartition théorique	C_1	C_2

la formule du χ^2 corrigé est alors :

$$\chi^2_{\text{corrigé}} = \frac{\left(|O_1 - C_1| - \frac{1}{2}\right)^2}{C_1} + \frac{\left(|O_2 - C_2| - \frac{1}{2}\right)^2}{C_2}$$

– Dans le cas de la comparaison de deux répartitions observées à 2 classes,

	Classe 1	Classe 2
Répartition observée	O_{11} (C_{11})	O_{12} (C_{12})
Répartition observée 2	O_{21} (C_{21})	O_{22} (C_{22})

la formule du χ^2 corrigé est alors :

$$\chi^2_{\text{corrigé}} = \frac{\left(|O_{11} - C_{11}| - \frac{1}{2}\right)^2}{C_{11}} + \frac{\left(|O_{12} - C_{12}| - \frac{1}{2}\right)^2}{C_{12}} + \frac{\left(|O_{21} - C_{21}| - \frac{1}{2}\right)^2}{C_{21}} + \frac{\left(|O_{22} - C_{22}| - \frac{1}{2}\right)^2}{C_{22}}$$

Soit $$\chi^2_{\text{corrigé}} = \sum_{i,j} \frac{\left(|O_{ij} - C_{ij}| - 0{,}5\right)^2}{C_{ij}}$$ avec $i = 1, 2$ (lignes), $j = 1, 2$ (colonnes)

• *Procédure à partir d'un exemple :*

On souhaite déterminer si les répartitions selon le sexe (féminin, masculin) de deux familles de 20 personnes sont équivalentes ou non. Les effectifs observés sont présentés dans le tableau 5 suivant, les effectifs calculés étant notés entre parenthèses.

	Femmes	Hommes	Total
Famille A	14	6	20
	(14,5)	(4,5)	
Famille B	17	3	20
	(14,5)	(4,5)	
Total	31	9	40

Tableau 5

Les hypothèses à tester sont :

H_0 : Les répartitions hommes/femmes sont équivalentes d'une famille à l'autre

H_1 : Les répartitions observées sont différentes

ou encore

H_0 : Il n'y a pas de liaison entre les caractères observés (sexe ratio et famille)

H_1 : Il y a une liaison entre les caractères qualitatifs étudiés.

Les conditions d'application du test du χ^2 n'étant pas vérifiées, puisque tous les effectifs calculés ne sont pas supérieurs ou égaux à 5, on utilise le χ^2 corrigé de Yates.

$$\chi^2_{\text{corrigé}} = \sum_{i,j} \frac{\left(|O_{ij} - C_{ij}| - 0,5\right)^2}{C_{ij}} \qquad \text{avec } i = 1, 2 \quad j = 1, 2$$

$$\chi^2_{corrigé} = \frac{(|14 - 14,5| - 0,5)^2}{14,5} + \frac{(|6 - 4,5| - 0,5)^2}{4,5} + \frac{(|17 - 14,5| - 0,5)^2}{14,5} + \frac{(|3 - 4,5| - 0,5)^2}{4,5}$$

$$\chi^2_{corrigé} = \frac{(0)^2}{14,5} + \frac{(1)^2}{4,5} + \frac{(2)^2}{14,5} + \frac{(1)^2}{4,5}$$

$$\chi^2_{corrigé} = \frac{1}{4,5} + \frac{4}{14,5} + \frac{1}{4,5}$$

$$\chi^2_{corrigé} = 0,44$$

On cherche alors, pour le risque α, la valeur du χ^2 tabulé dans la table du χ^2 pour un nombre de degrés de liberté égal au nombre de ligne moins un multiplié par le nombre de colonnes moins un (soit toujours ddl = 1).
Car ddl = (2 - 1) x (2 - 1) = 1

Alors pour $\alpha = 5$ % et un ddl = 1, la valeur du χ^2 tabulé est $\chi^2_T = 3{,}841$.

Donc $\chi^2_{corrigé} < \chi^2_T$ alors on rejette l'hypothèse alternative H_1 et on conclut à une différence non significative entre les répartitions observées hommes/femmes selon les familles A ou B. Ainsi il n'y a pas de liaison entre les deux caractères qualitatifs étudiés, le sexe est indépendant de l'appartenance à la famille A ou B. Les répartitions hommes/femmes sont équivalentes d'une famille à l'autre.

2. Données appariées - χ^2 de Mac Nemar

Rôle du test :

Le test du χ^2 de Mac Nemar permet de comparer deux répartitions observées dans le cas de **deux séries appariées**.

• Procédure à partir d'un exemple :

Imaginons qu'un psychologue souhaite comparer deux thérapies A et B en terme d'efficacité ressentie par 100 malades qui ont été leur propre témoin. Ces malades ont tous été soumis successivement aux deux thérapies dans un ordre tiré au hasard.

Les hypothèses à tester sont :

$\begin{cases} H_0 : \text{L'efficacité des thérapies A et B est équivalente} \\ H_1 : \text{L'efficacité des thérapies A et B est différente} \end{cases}$

Le résultat de chaque thérapie a été noté + ou -, de sorte que l'on obtient finalement 100 couples de réponses. Quatre couples de réponses sont possibles : (-, -), (-, +), (+, -) ou (+, +). Le couple (-, +) signifie que la thérapie B est ressentie comme efficace et la thérapie A comme non efficace. Les données observées sont résumées dans le tableau 6 suivant.

Thérapies A	Thérapies B	Nombre de malades
-	-	40
-	+	20
+	-	10
+	+	30
Total		100

Tableau 6

Les données étant appariées, le test du χ^2 classique n'est pas adapté dans ce cas. Le test du χ^2 peut être utilisé, mais il est moins puissant puisque l'on ne tient pas compte de toute l'information disponible. C'est pourquoi on utilise plutôt le χ^2 de Mac Nemar.

On ne tient pas compte des couples de réponses concordantes, c'est-à-dire des 40 couples (-, -) et de 30 couples (+, +), qui n'apportent aucune information sur la meilleure efficacité des thérapies A ou B. En revanche, on considère les 30 couples de réponses discordantes, soit les 20 couples de réponses (-, +) et les 10 couples (+, -). L'hypothèse d'équivalence des deux thérapies suppose que l'on observe autant de couples de réponses (-, +) que de couples (+, -). Il s'agit alors de comparer le pourcentage de réponses (-, +) sur l'ensemble des réponses discordantes, c'est-à-dire : $\frac{20}{30}$, à $\frac{1}{2}$

(ce qui revient à comparer le pourcentage de réponses (+, -), soit $\frac{10}{30}$ à $\frac{1}{2}$).

Si on note les nombres de couples de réponses discordantes a et b,

soit a : le nombre de couples (-, +)
et b : le nombre de couples (+, -)

cela revient à comparer $\frac{a}{a+b}$ à $\frac{1}{2}$

ou encore à comparer a à $\frac{a+b}{2}$

par la formule du χ^2 de Mac Nemar : $\chi^2 = \frac{\left(a - \frac{a+b}{2}\right)^2}{(a+b)\frac{1}{2}\frac{1}{2}}$

soit $$\chi^2 = \frac{(a-b)^2}{a+b}$$

que l'on compare à son tour à la valeur du χ^2 lu dans la table du χ^2 à un degré de liberté.

On a calculé $\chi^2_C = 3{,}33$ et lu pour $\alpha = 5\ \%$ $\chi^2_T = 3{,}841$;
donc $\chi^2_C < \chi^2_T$ alors on rejette l'hypothèse alternative H_1 et on conclut à une différence non significative entre les séries appariées. Ainsi les thérapies A et B sont équivalentes en terme d'efficacité ressentie par les 100 malades observés.

Si $\chi^2_C < \chi^2_T$ alors on rejette l'hypothèse alternative H_1 et on conclut à une différence non significative entre les répartitions observées, c'est-à-dire à une équivalence entre les deux séries appariées.
Si $\chi^2_C \geq \chi^2_T$ alors on rejette l'hypothèse nulle et on conclut à une différence significative entre les répartitions observées, c'est-à-dire à une différence entre les deux séries appariées. Puis, la valeur de α (lue dans la table du χ^2) correspondant au χ^2_T le plus proche et inférieur au χ^2_C fixe le degré de signification de la différence.

■ ***Conditions d'application :***

Cette méthode n'est utilisable que si $\frac{a+b}{2} \geq 5$ c'est-à-dire si l'on observe au moins 10 couples de réponses discordantes.

Si ces conditions d'application ne sont pas vérifiées, c'est-à-dire si $2 < \frac{a+b}{2} < 5$ on peut utiliser le χ^2 corrigé :

$$\chi^2_{\text{corrigé}} = \frac{(|a-b|-1)^2}{a+b}$$

3. Récapitulatif

(en page 165)

Récapitulatif des tests à utiliser dans le cas de variables qualitatives (classées)

Tests	Rôle du test	Type de variables	Population	Conditions d'applications	Formule	Table	Conclusion
χ^2	- Comparaison d'une répartition observée à une répartition théorique - Comparaison de plusieurs répartitions observées - Etude de la liaison entre deux variables classées	Qualitatives	Un, deux ou plus de deux groupes indépendants	Tous les effectifs calculés ou théoriques sont supérieurs à 5 ($Ci \geq 5$)	$\chi^2_c = \sum_{i,j} \frac{(O_{ij} - C_{ij})^2}{C_{ij}}$	χ^2	ddl = (c - 1)x(l - 1) si $\chi^2_c \geq \chi^2_T$ alors on rejette H_0 au seuil α
χ^2 corrigé de Yates	- Comparaison de deux répartitions observées - Etude de la liaison entre 2 variables	Qualitatives à 2 classes	Deux groupes indépendants	Tous les effectifs calculés ou théoriques sont compris entre 2 et 5 ($2 < Ci < 5$)	$\chi^2_Y = \sum_{i,j} \frac{\left(\lvert O_{ij} - C_{ij} \rvert - \frac{1}{2}\right)^2}{C_{ij}}$	χ^2	ddl = 1 si $\chi^2_Y \geq \chi^2_T$ alors on rejette H_0 au seuil α
χ^2 de Mac Nemar	- Comparaison de deux répartitions observées dans le cas apparié	Qualitatives à 2 classes	Deux séries appariées	$\frac{a+b}{2} \geq 5$	$\chi^2_c = \frac{(a-b)^2}{a+b}$	χ^2	ddl = 1 si $\chi^2_c \geq \chi^2_T$ alors on rejette H_0 au seuil α
				$2 < \frac{a+b}{2} < 5$	si $2 < \frac{a+b}{2} < 5$ $\chi^2_{corrigé} = \frac{(\lvert a-b \rvert - 1)^2}{a+b}$		

H_0 : distributions équivalentes
H_1 : distributions différentes

Tableau 7

4. Exercices

4.1. Exercices corrigés (solutions pages 169 à 172)

• *Exercice 1 :*

Une étudiante en psychologie étudie, dans le cadre d'une thèse en psychologie de la santé, la répartition, selon le sexe, de 110 patients présentant une pathologie P. Elle dénombre 45 hommes et veut savoir si cet effectif observé est conforme au modèle d'égalité d'atteinte selon le sexe.
Indiquer la méthode qu'elle doit utiliser pour répondre à cette question.
Résoudre le problème au risque habituel de 5 %.

• *Exercice 2 :*

Les nombres d'enfants de deux échantillons de 100 femmes ménopausées, du nord et du sud de la France, ont été observés. Les résultats sont présentés dans le tableau suivant :

	Nombre d'enfants			
	0	**1 ou 2**	**3 ou 4**	**plus de 4**
Femmes du nord	20	54	15	11
Femmes du sud	15	50	18	17

La répartition du nombre d'enfants est-elle la même pour les femmes du nord et du sud ?
1. Indiquer la méthodologie à employer afin de pouvoir répondre à cette question.
2. Indiquer le nom du test statistique à réaliser dans de bonnes conditions méthodologiques pour répondre à la question. Poser dans ce cas les hypothèses, et appliquer le test approprié.

• *Exercice 3 :*

Un groupe de psychiatres a observé que, parmi les 800 patients examinés dans leur consultation en dispensaire, il y avait 185 ouvriers, 250 employés et cadres moyens, 250 cadres supérieurs ou membres de professions libérales et 115 patients d'autres catégories professionnelles.
Dans le même département, les données INSEE pour les mêmes catégories professionnelles sont respectivement pour 100 personnes : 36, 30, 20 et 14.
1. La répartition selon la catégorie professionnelle dans la consultation reflète-t-elle celle de la population du département ?
2. Sinon, comment l'expliquer ?

4.2. Exercices non corrigés

• *Exercice 4 :*

On souhaite comparer les temps de réponse d'un groupe de sujets répondant à deux tests psychotechniques T1 et T2 successifs. Pour cela, on mesure successivement le temps de réponse au test T1, puis le lendemain, le temps de réponse au test T2, pour chaque individu d'un échantillon randomisé de 60 personnes issues de la population d'un village du sud-ouest de la France. L'ordre des tests effectués a été tiré au sort pour chaque individu, afin de supprimer l'impact éventuel de "l'effet d'ordre" sur les résultats et les temps de réponse. Les temps de réponse sont plus ou moins longs selon les deux tests.
On observe :
- Chez 22 sujets, un temps de réponse au test T1 moins long qu'au test T2.
- Chez 35 sujets, un temps de réponse au test T1 plus long qu'au test T2.

Il y a donc eu, au total, 57 observations de temps de réponse discordantes. Peut-on dire que les temps de réponse sont différents selon le test T1 ou T2 ? Justifiez de façon détaillée votre réponse.

• *Exercice 5 :*

1. On étudie un échantillon de 100 sujets que l'on soumet à un test A dont le résultat final est 0 ou 1. On observe que 40 sujets ont obtenu un résultat de 1, et les 60 autres ont obtenu 0.
On sait qu'en règle générale, 50 % des sujets ont un résultat de 1, et 50 % ont un résultat de 0. On souhaite savoir si l'échantillon observé obtient des résultats conformes à ceux de la population générale. Effectuer le test approprié.

2. On étudie un autre groupe de 150 sujets, que l'on soumet au même test A que précédemment. On observe alors que 80 sujets ont obtenu un résultat de 0. Ce deuxième échantillon obtient-il des résultats conformes à ceux du premier échantillon ?

3. Un test B du même type, mais dont les résultats possibles sont 0, 1 ou 2, est également proposé au premier échantillon de sujets.

30 sujets obtiennent 0,
40 sujets obtiennent 1,
30 sujets obtiennent 2.

En règle générale, dans la région :

20 % des sujets obtiennent 0,
40 % des sujets obtiennent 1,
40 % des sujets obtiennent 2.

Les résultats obtenus par l'échantillon sont-ils conformes à ceux obtenus dans la région ?

4. Le deuxième groupe de sujets est soumis au test B, et les résultats sont les suivants :

30 sujets obtiennent 0,
50 sujets obtiennent 1,
70 sujets obtiennent 2.

Les résultats obtenus par l'échantillon sont-ils conformes à ceux obtenus dans le région ?

Les résultats obtenus par le deuxième groupe sont-ils conformes à ceux obtenus par le premier groupe ?

5. Un test C dont le résultat varie entre 0 et 100 est proposé au premier échantillon. L'ensemble des sujets obtiennent en moyenne un score de 53,4 ; de variance 22. Un an plus tard, on soumet ce même échantillon au test C, et ils obtiennent en moyenne un score de 58,7 ; de variance 20. Ces deux moyennes diffèrent-elles ?

6. Le deuxième échantillon est soumis au test C et obtient en moyenne un score de 60,2 ; de variance 18. Cette moyenne diffère-t-elle de celle obtenue par le premier échantillon la première année ? Même question pour la deuxième année.

• *Exercice 6 :*

Un psychologue fait l'hypothèse que certaines difficultés de langage écrit chez l'enfant sont en relation avec la latéralisation.

Sur un échantillon de 108 enfants, ce psychologue recueille les données suivantes :

Latéralisation	nombre d'enfants ayant des difficultés de langage écrit	nombre d'enfants n'ayant aucune difficulté de langage écrit
droitier	15	30
ambidextre	18	10
gaucher	20	15

En examinant ces résultats,

1. Formuler les hypothèses avec précision.

2. Eprouver ces hypothèses à l'aide du test statistique approprié.

3. Peut-on conclure que la latéralisation a une incidence sur les difficultés de langage écrit ?

• *Exercice 7 :*

Une enquête se propose de comparer l'utilisation ou non de préservatifs chez deux groupes de personnes classés selon l'âge. Les résultats sont présentés dans le tableau suivant :

	Personnes de moins de 30 ans	Personnes de 30 ans et plus
Préservatifs	520	410
Pas de préservatif	380	490

1. Les moins de 30 ans et les 30 ans et plus utilisent-ils des préservatifs avec la même fréquence ? (au seuil $\alpha = 5\ \%$)
2. Interpréter ces résultats.

4.3. Solutions

• *Correction de l'exercice 1 :*

Le modèle d'égalité d'atteinte selon le sexe correspond à une répartition théorique d'atteinte par la pathologie P de 50 % de femmes et 50 % d'hommes. La question est alors de savoir si la répartition observée est équivalente ou non à cette répartition théorique.

Les hypothèses à tester sont donc :

H_0 : la répartition observée est équivalente à la répartition théorique ; ou, les effectifs observés sont conformes au modèle d'égalité d'atteinte selon le sexe.

contre

H_1 : la répartition observée est différente de la répartition théorique ; les effectifs observés ne sont pas conformes au modèle d'égalité d'atteinte selon le sexe.

La variable étudiée (atteinte par la pathologie P selon le sexe) étant qualitative, le test permettant de tester ces hypothèses est le test du χ^2 et le tableau de calcul du χ^2 est le suivant :

	Femmes présentant la pathologie P	Hommes présentant la pathologie P	Total
Effectifs observés (O_i)	65	45	110
Répartition théorique	50	50	100
Effectifs théoriques (C_i)	55	55	110
$O_i - C_i$	10	-10	0 (toujours)

Ici nous avons ramené le total des effectifs théoriques au total des effectifs observés afin de rendre les répartitions marginales égales (bien entendu, les

effectifs observés ne doivent en aucun cas être modifiés, sinon par de nouvelles enquêtes basées sur des échantillons plus larges).
Tous les effectifs théoriques étant supérieurs à 5, les conditions d'application sont vérifiées, ce qui nous permet de déterminer le χ^2 calculé.

$$\chi^2_C = \sum_i \frac{(o_i - c_i)^2}{c_i}$$

$$\chi^2_C = \frac{(10)^2}{55} + \frac{(-10)^2}{55} = 3{,}64$$

Au risque $\alpha = 5\ \%$, avec un nombre de degrés de liberté de $(2-1) = 1$, le paramètre "tabulé" (lu dans la table du χ^2) est $\chi^2_T = 3{,}841$.

Donc $\chi^2_C < \chi^2_T$, soit $\chi^2_C \in IP_{(95\%)}[0\,;3{,}841]$, alors on ne rejette pas l'hypothèse nulle.
Au risque $\alpha = 5\ \%$, on conclut alors que la différence entre la répartition observée et la répartition théorique n'est pas significative, ou encore que la répartition observée de femmes et d'hommes atteints par la pathologie P est conforme au modèle d'égalité d'atteinte selon le sexe.

• *Correction de l'exercice 2 :*

1. Pour répondre à la question "les femmes du sud et du nord de la France ont-elles eu autant d'enfants ?", il faudrait que les deux échantillons aient été constitués par tirage au sort dans les deux populations de femmes. C'est-à-dire que les deux échantillons soient représentatifs des deux populations de référence. Dans le cas contraire, on pourrait montrer une différence entre ces deux répartitions observées, ou un lien entre les deux caractères étudiés (nombre d'enfants et situation géographique), uniquement en spécifiant que la signification ou non de la différence ne concerne que les femmes des deux échantillons observés, et en précisant le mode de constitution des échantillons.
2. Le test à effectuer pour répondre à la question posée est un test du χ^2.
Les hypothèses à tester sont alors :

H_0 : Les deux répartitions observées sont équivalentes ; ou, les femmes du sud et du nord de la France ont autant d'enfants.

contre

H_1 : Les deux répartitions observées sont différentes ; ou, les femmes du sud et du nord de la France n'ont pas autant d'enfants.

Les variables femmes du nord et du sud de la France, et nombre d'enfants (variable classée) étant qualitatives, le test du χ^2 peut permettre de tester ces hypothèses selon la vérification des conditions d'application.
Le tableau de calcul du χ^2 est le suivant (les effectifs calculés étant entre parenthèses) :

	Nombre d'enfants				
	0	1 ou 2	3 ou 4	plus de 4	Total
Femmes du nord	20	54	15	11	100
	(17,5)	(52)	(16,5)	(14)	
Femmes du sud	15	50	18	17	100
	(17,5)	(52)	(16,5)	(14)	
Total	35	104	33	28	200

Tous les effectifs théoriques étant supérieurs à 5, les conditions d'application sont vérifiées, ce qui nous permet de déterminer le χ^2 calculé.
Le paramètre à calculer est alors :

$$\chi^2_C = \sum_{ij} \frac{(o_{ij} - c_{ij})^2}{c_{ij}}$$

$$\chi^2_C = \frac{(20-17,5)^2}{17,5} + \frac{(54-52)^2}{52} + \frac{(15-16,5)^2}{16,5} + \frac{(11-14)^2}{14} + \frac{(15-17,5)^2}{17,5} + \frac{(50-52)^2}{52}$$
$$+ \frac{(18-16,5)^2}{16,5} + \frac{(17-14)^2}{14} = 2,43$$

Au risque $\alpha = 5\ \%$, avec un nombre de degré de liberté de (2-1) x (4-1) = 3, le paramètre "tabulé" est $\chi^2_T = 7,815$.
Donc $\chi^2_C < \chi^2_T$, soit $\chi^2_C \in \mathbb{P}_{(95\%)}[0\ ;\ 7,815]$, alors on ne rejette pas l'hypothèse nulle.
Au risque $\alpha = 5\ \%$, on conclut alors que la différence entre les deux répartitions observées n'est pas significative, ou encore que la liaison entre les caractères "nombre d'enfants" et "situation géographique nord sud" n'est pas significative. Ainsi, les femmes du sud et du nord de la France ont eu autant d'enfants.

• *Correction de l'exercice 3 :*

1. Afin de déterminer si la répartition selon la catégorie professionnelle dans la consultation reflète ou non celle de la population du département, il faut tester les hypothèses suivantes :

- H_0 : la répartition observée est équivalente à la répartition théorique ; ou, la catégorie professionnelle dans la consultation reflète celle de la population du département
- H_1 : la répartition observée est différente de la répartition théorique ; la catégorie professionnelle dans la consultation ne reflète pas celle de la population du département.

Les variables étudiées étant qualitatives, le test permettant de tester ces hypothèses est le test du χ^2.

Le tableau de calcul du χ^2 est le suivant :

	Ouvriers	Employés	Cadres sup.	Autres	Total
Effectifs observés (Oi)	185	250	250	115	800
Répartition théorique	36	30	20	14	100
Effectifs théoriques (Ci)	288	240	160	112	800
Oi - Ci	-103	10	90	3	0 (toujours)

Tous les effectifs théoriques étant supérieurs à 5, les conditions d'application sont vérifiées, ce qui nous permet de déterminer le χ^2 calculé.

$$\chi^2_C = \sum_i \frac{(o_i - c_i)^2}{c_i}$$

$$\chi^2_C = \frac{(-103)^2}{288} + \frac{(10)^2}{240} + \frac{(90)^2}{160} + \frac{(3)^2}{112} = 87{,}96$$

Au risque $\alpha = 5\ \%$, avec un nombre de degré de liberté de (4-1) = 3, le paramètre "tabulé" (lu dans la table du χ^2) est $\chi^2_T = 7{,}815$.

Donc $\chi^2_C > \chi^2_T$, soit $\chi^2_C \notin IP_{(95\%)}[0\,;7{,}815]$, alors on rejette l'hypothèse nulle. On conclut alors que la différence entre la répartition observée et la répartition théorique est significative ($p < 0{,}001$) ; c'est-à-dire que la répartition selon la catégorie professionnelle dans la consultation ne reflète pas celle de la population du département.

2. Les troubles psychiatriques n'affectent pas les catégories socioprofessionnelles dans les mêmes proportions que dans la consultation observée. On pourrait penser que l'accès aux soins diffère d'une catégorie à l'autre, ou encore que certaines catégories consultent davantage en milieu hospitalier que d'autres.

CHAPITRE VII

Tests des différences entre variables quantitatives dans le cas de deux groupes

1. Variables quantitatives dans le cas de deux groupes avec données non appariées *175*

1.1. Comparaison d'une moyenne observée à une moyenne théorique ($n \geq 30$) - test ε ou test z *175*

1.1.1. Rôle du test ε ou test z *175*

1.1.2. Procédure du test *175*

1.1.3. Conditions d'application du test *176*

1.2. Comparaison d'une moyenne observée à une moyenne théorique ($n < 30$) - test t *179*

1.2.1. Rôle du test t *179*

1.2.2. Procédure du test *179*

1.2.3. Conditions d'application *180*

1.3. Comparaison de deux moyennes observées (avec n_A et $n_B \geq 30$) - test ε ou test z *182*

1.3.1. Rôle du test ε ou test z *182*

1.3.2. Procédure du test *182*

1.3.3. Conditions d'application *183*

1.4. Comparaison de deux moyennes observées (avec n_A ou $n_B < 30$ et distributions normales de même variance) - test t *185*

1.4.1. Rôle du test t *185*

1.4.2. Procédure du test *185*

1.4.3. Conditions d'application *186*

1.5. Comparaison de deux lois observées (avec $n < 30$ et distribution non normale) - test de Mann et Whitney *188*

1.5.1. Rôle du test de Mann et Whitney *188*

1.5.2. Procédure générale du test de Mann et Whitney *188*

1.5.3. Test de Mann et Whitney si $n_1 \geq 9$ et $n_2 \geq 9$, échantillons sans nœud *191*

1.5.4. Test de Mann et Whitney si $n_1 \geq 9$ et $n_2 \geq 9$, échantillons avec nœud *194*

1.5.5. Test de Mann et Whitney
si $n_1 \leq 8$ et $n_2 \leq 8$, et échantillons sans nœud *199*
1.5.6. Test de Mann et Whitney
si $n_1 \leq 8$ et $9 \leq n_2 \leq 20$ et échantillon sans nœud *199*
1.5.7. Test de Mann et Whitney si $n_1 \leq 8$ et $n_2 > 20$,
et échantillons sans nœud ou avec nœud *200*
1.5.8. Test de Mann et Whitney
si $n_1 \leq 8$ et $n_2 \leq 8$, et échantillons avec nœud *200*
1.6. Comparaison de deux variances *201*
1.6.1. Rôle du test de comparaison de deux variances *201*
1.6.2. Procédure du test *202*
1.6.3. Lecture des tables de F *203*
1.6.4. Procédure résumée *203*
1.6.5. Conditions d'application *204*

2. Variables quantitatives dans le cas de deux groupes avec données appariées *204*
2.1. Comparaison de deux moyennes observées (avec $n \geq 30$ et distribution normale) - test ε (ou test z) *204*
2.1.1. Rôle du test *204*
2.1.2. Procédure du test *204*
2.1.3. Conditions d'application *206*
2.2. Comparaison de deux moyennes observées (avec $n < 30$ et distribution normale) - test t *208*
2.2.1. Rôle du test *208*
2.2.2. Procédure du test *208*
2.2.3. Conditions d'application *209*
2.3. Comparaison de deux moyennes observées - test de Wilcoxon *211*
2.3.1. Rôle du test *211*
2.3.2. Procédure du test *211*
2.3.3. Test de Wilcoxon
si échantillon de petite taille $n \leq 20$, sans nœud *213*
2.3.4. Test de Wilcoxon
si échantillon de petite taille $n \leq 20$, avec nœud *214*
2.3.5. Test de Wilcoxon
si échantillon de grande taille $n > 20$, sans nœud *214*
2.3.6. Test de Wilcoxon
si échantillon de grande taille $n > 20$, avec nœud *215*

3. Récapitulatif *215*

CHAPITRE VII

Tests des différences entre variables quantitatives dans le cas de deux groupes

1. Variables quantitatives dans le cas de deux groupes avec données non appariées

1.1. Comparaison d'une moyenne observée à une moyenne théorique ($n \geq 30$) - test ε ou test z

1.1.1. Rôle du test ε ou test z

Le test ε permet de **comparer une moyenne observée à une moyenne théorique**, dans le cas de grands échantillons de taille $n \geq 30$. Il s'agit de déterminer si l'échantillon observé de moyenne m (concernant la variable quantitative étudiée) peut être considéré comme issu de la population de référence de moyenne théorique μ.

1.1.2. Procédure du test

Les hypothèses à tester pour comparer une moyenne observée à une moyenne théorique sont :

H_0 : la moyenne observée est équivalente à la moyenne théorique,

contre
- H_1 : la moyenne observée est différente de la moyenne théorique,
- ou H_1' : la moyenne observée est inférieure à la moyenne théorique,
- ou H_1'' : la moyenne observée est supérieure à la moyenne théorique,

que l'on peut formuler ainsi :

$$\begin{cases} H_0 : m = \mu \\ \text{contre} \begin{cases} H_1 : m \neq \mu \\ \text{ou } H_1' : m < \mu \\ \text{ou } H_1'' : m > \mu \end{cases} \end{cases}$$

■ ***Estimation de l'écart-réduit calculé :***

La comparaison d'une moyenne m, observée sur n individus, à une valeur théorique μ, est basée sur l'estimation de l'écart-réduit calculé (c'est-à-dire ε_c) :

$$\varepsilon_c = \frac{m - \mu}{\frac{s}{\sqrt{n}}}$$

avec s correspondant à **l'écart-type de la variable quantitative étudiée**, estimé sur l'échantillon de taille n

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - m)^2}{n - 1}$$

▩ ***Remarque :***

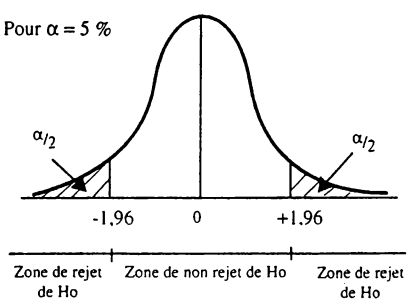
Quand σ^2 est connu, on utilise sa valeur dans le calcul de ε_c.

1.1.3. Conditions d'application du test

Le test ε de comparaison d'une moyenne observée à une moyenne théorique n'est utilisable que si **la taille de l'échantillon est supérieure ou égale à 30** ($n \geq 30$).

➤ ***Dans le cas d'un test bilatéral (H_1)***

Au seuil $\alpha = 5\ \%$, ε lu dans la table de la loi normale correspond à $\varepsilon_t = 1{,}96$



Si $|\varepsilon_c| < 1{,}96$, c'est-à-dire si ε se situe dans la zone de non rejet de H_0, alors on ne rejette pas l'hypothèse H_0. Et on conclut à une différence non significative entre la moyenne observée et la moyenne théorique au risque $\alpha = 5\ \%$, avec ε tabulé, ($\varepsilon_t = 1{,}96$), lu dans la table de la loi normale. On conclut alors, concernant la variable quantitative étudiée, que l'échantillon observé est issu de la population de référence de moyenne théorique μ.

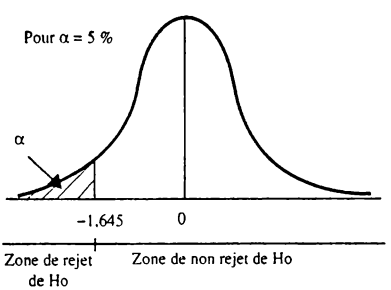
Si $|\varepsilon_c| \geq 1{,}96$, c'est-à-dire si ε se situe dans la zone de rejet de H_0, alors on rejette l'hypothèse H_0. Et on conclut à une différence significative entre la moyenne observée et la moyenne théorique, au risque $\alpha = 5\ \%$.

Le risque α', lu dans la table de la loi normale, correspondant à la valeur la plus proche de ε_c mais inférieure (en valeur absolue), **fixe le degré de signification de la différence** (on le note : $p < \alpha'$). Le degré de signification de la différence est le risque α' jusqu'auquel on conclut encore à une différence significative, c'est **le risque minimum de rejeter H_0 à tort**.

• Exemple :

Si $\varepsilon_c = 2{,}3$, il se situe dans la zone de rejet de H_0. Si on rejette H_0, on ne rejette pas H_1, et on conclut à une différence significative entre la moyenne observée et la moyenne théorique. Puis on recherche dans la table de la loi normale la valeur de α' correspondant à la valeur la plus proche de ε_c mais inférieure, soit pour $\varepsilon_t = 2{,}17$ on a $\alpha' = 0{,}03 = 3\ \%$. Ainsi $\alpha' = 0{,}03$ fixe le degré de signification de la différence (on le note : $p < 0{,}03$).

➤ ***Dans le cas d'un test unilatéral à gauche (H'_1)***



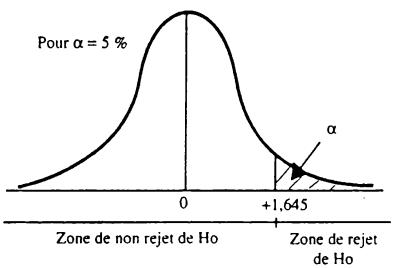
Si $\varepsilon_c > -1{,}645$, c'est-à-dire si ε se situe dans la zone de non rejet de H_0, alors non rejet de l'hypothèse H_0. Et on conclut à une inégalité à gauche non significative entre la moyenne observée et la moyenne théorique au risque $\alpha = 5\ \%$, avec ε tabulé, ($\varepsilon_t = -1{,}645$), lu dans la table de la

loi normale (pour α_t = 10 %). On conclut alors, concernant la variable quantitative étudiée, que l'échantillon observé est issu de la population de référence de moyenne théorique μ.

Si $\varepsilon_c \leq -1{,}645$, c'est-à-dire si ε se situe dans la zone de rejet de H_0, alors on rejette l'hypothèse H_0. Et on conclut à une inégalité à gauche significative entre la moyenne observée et la moyenne théorique, au risque α = 5 %. Ainsi, la moyenne observée est significativement inférieure à la moyenne théorique.

Le risque α', lu dans la table de la loi normale, correspondant à la valeur la plus proche de ε_c mais inférieure (en valeur absolue), **fixe le degré de signification de l'inégalité** (on le note : $p < \alpha'$).

➤ *Dans le cas d'un test unilatéral à droite (H'_1)*



Si $\varepsilon_c < +1{,}645$, c'est-à-dire si ε se situe dans la zone de non rejet de H_0, alors on ne rejette pas l'hypothèse H_0. Et on conclut à une inégalité à droite non significative entre la moyenne observée et la moyenne théorique au risque α = 5 %, avec ε tabulé, ($\varepsilon_t = +1{,}645$), lu dans la table de la loi normale (pour α_t = 10 %). On conclut alors, concernant la variable quantitative étudiée, que l'échantillon observé est issu de la population de référence de moyenne théorique μ.

Si $\varepsilon_c \geq +1{,}645$, c'est-à-dire si ε se situe dans la zone de rejet de H_0, alors on rejette l'hypothèse H_0. Et on conclut à une inégalité à droite significative entre la moyenne observée et la moyenne théorique, au risque α = 5 %. Ainsi, la moyenne observée est significativement supérieure à la moyenne théorique.

Le risque α', lu dans la table de la loi normale, correspondant à la valeur la plus proche de ε_c mais inférieure (en valeur absolue), **fixe le degré de signification de l'inégalité** (on le note : $p < \alpha'$).

1.2. Comparaison d'une moyenne observée à une moyenne théorique (n < 30) - test t

1.2.1. Rôle du test t

Le test t permet de **comparer une moyenne observée à une moyenne théorique, dans le cas de petits échantillons de taille n < 30**, si le caractère quantitatif étudié est distribué selon une **loi normale**. Il s'agit de déterminer si l'échantillon observé de moyenne m (concernant la variable quantitative étudiée) peut être considéré comme issu de la population de référence de moyenne théorique μ.

1.2.2. Procédure du test

Les hypothèses à tester pour comparer une moyenne observée à une moyenne théorique sont :

H_0 : la moyenne observée est équivalente à la moyenne théorique,

contre
- H_1 : la moyenne observée est différente de la moyenne théorique,
- ou H_1' : la moyenne observée est inférieure à la moyenne théorique,
- ou H_1'' : la moyenne observée est supérieure à la moyenne théorique,

que l'on peut formuler ainsi :

$H_0 : m = \mu$

contre
- $H_1 : m \neq \mu$
- ou $H_1' : m < \mu$
- ou $H_1'' : m > \mu$

■ *Estimation du paramètre t_c :*

La comparaison d'une moyenne m, observée sur n individus, à une valeur théorique μ, est basée sur l'estimation du paramètre t_c calculé :

$$t_c = \frac{m - \mu}{\frac{s}{\sqrt{n}}}$$

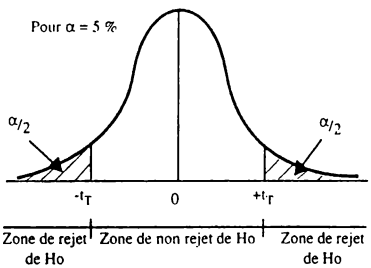
avec s correspondant à l'écart-type de la variable quantitative étudiée, estimé sur l'échantillon de taille n :

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - m)^2}{n - 1}$$

1.2.3. Conditions d'application

Le test t de comparaison d'une moyenne observée à une moyenne théorique n'est utilisable que **si le caractère quantitatif étudié est distribué selon une loi normale.**

➤ *Dans le cas d'un test bilatéral (H_1)*



Si $|t_c| < t_T$, c'est-à-dire si t_c se situe dans la zone de non rejet de H_0, alors on ne rejette pas l'hypothèse H_0 et on conclut à une différence non significative entre la moyenne observée et la moyenne théorique au risque $\alpha = 5$ %. La valeur de t tabulé, t_T, est lue dans la table de t au risque α choisi pour un degré de liberté égal à n - 1. On conclut alors, concernant la variable quantitative étudiée, que l'échantillon observé est issu de la population de référence de moyenne théorique μ.

Si $|t_c| \geq t_T$, c'est-à-dire si t_c se situe dans la zone de rejet de H_0, alors on rejette l'hypothèse H_0 et on conclut à une différence significative entre la moyenne observée et la moyenne théorique, au risque $\alpha = 5$ %.

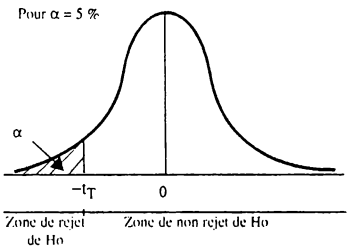
Le risque α', lu dans la table de t avec un nombre de degré de liberté égal à n - 1, correspondant à la valeur la plus proche de t_c mais inférieure (en valeur absolue), fixe le degré de signification de la différence (on le note : $p < \alpha'$). Le degré de signification de la différence est le risque α' jusqu'auquel on conclut encore à une différence significative, c'est **le risque minimum de rejeter H_0 à tort**.

• *Exemple :*

Si $t_c = 5$, avec un ddl = 3, pour $\alpha = 5$ % on a $t_T = 3{,}182$, alors t_c se situe dans la zone de rejet de H_0. Si on rejette H_0 (on ne rejette pas H_1) on conclut à une différence significative entre la moyenne observée et la moyenne théorique. Puis on recherche dans la table de t la valeur de α' correspondant à la valeur la plus proche de t_c mais inférieure en valeur absolue. Soit avec un ddl = 3, pour $t_T = 4{,}541$ on a $\alpha' = 0{,}02 = 2$ %. Ainsi $\alpha' = 0{,}02$ fixe le degré de signification de la différence. En revanche, avec un ddl = 3, pour $t_T = 5{,}841$ on a

$\alpha' = 0{,}01 = 1\ \%$, mais t_c étant inférieur à t_T on ne conclut plus à une différence significative ; c'est pourquoi le degré de signification de la différence est $p < 0{,}02$ (selon la table de t jointe en annexe p se situe entre 0,01 et 0,02).

➤ *Dans le cas d'un test unilatéral à gauche (H'_1)*

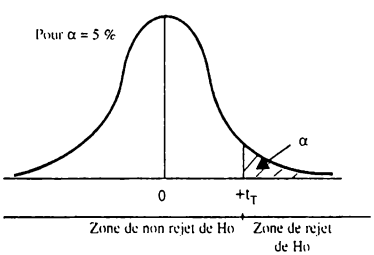


Si $t_c > -t_T$, c'est-à-dire si t_c se situe dans la zone de non rejet de H_0, alors on ne rejette pas l'hypothèse H_0. Et on conclut que la moyenne observée n'est pas plus petite que la moyenne théorique au risque $\alpha = 5\ \%$, avec t tabulé, t_T, lu dans la table de t avec un nombre de degré de liberté égal à n - 1, pour $\alpha_T = 10\ \%$. On conclut alors, concernant la variable quantitative étudiée, que l'échantillon observé est issu de la population de référence de moyenne théorique μ.

Si $t_c \leq -t_T$, c'est-à-dire si t_c se situe dans la zone de rejet de H_0, alors on rejette l'hypothèse H_0. Et on conclut à une inégalité à gauche significative entre la moyenne observée et la moyenne théorique au risque $\alpha = 5\ \%$, avec t tabulé, t_T, lu dans la table de t avec un nombre de degré de liberté égal à n - 1, pour $\alpha_T = 10\ \%$. Ainsi, la moyenne observée est significativement inférieure à la moyenne théorique.

Le risque α', lu dans la table de t avec ddl = n - 1, correspondant à la valeur la plus proche de t_c mais inférieure (en valeur absolue), **fixe le degré de signification de l'inégalité** (on le note : $p < \alpha'$).

➤ *Dans le cas d'un test unilatéral à droite (H'_1)*



Si $t_c < +t_T$, c'est-à-dire si t_c se situe dans la zone de non rejet de H_0, alors on ne rejette pas l'hypothèse H_0. Et on conclut à une inégalité à droite non significative entre la moyenne observée et la moyenne théorique au risque $\alpha = 5$ %, avec t tabulé, t_T, lu dans la table de t avec un nombre de degré de liberté égal à n - 1, pour $\alpha_T = 10$ %. On conclut alors, concernant la variable quantitative étudiée, que l'échantillon observé est issu de la population de référence de moyenne théorique μ.

Si $t_c \geq +t_T$, c'est-à-dire si t_c se situe dans la zone de rejet de H_0, alors on rejette l'hypothèse H_0. Et on conclut à une inégalité à droite significative entre la moyenne observée et la moyenne théorique, au risque $\alpha = 5$ %, avec t tabulé, t_T, lu dans la table de t avec un nombre de degré de liberté égal à n - 1, pour $\alpha_T = 10$ %. Ainsi, la moyenne observée est significativement supérieure à la moyenne théorique.

Le risque α', lu dans la table de t avec ddl = n - 1, correspondant à la valeur la plus proche de t_c mais inférieure (en valeur absolue), **fixe le degré de signification de l'inégalité** (on le note : $p < \alpha'$).

1.3. Comparaison de deux moyennes observées (avec n_A et $n_B \geq 30$) - test ε ou test z

1.3.1. Rôle du test ε ou test z

Le test ε permet de comparer **deux moyennes observées m_A et m_B**, dans le cas de **grands échantillons** de taille respective n_A et n_B supérieures ou égales à 30. Il s'agit de déterminer si les deux échantillons observés de moyenne m_A et m_B (concernant la variable quantitative étudiée) peuvent être considérés comme issus de la même population de référence de moyenne théorique commune $\mu = \mu_A = \mu_B$.

1.3.2. Procédure du test

Les hypothèses à tester pour comparer deux moyennes observées sont :

H_0 : les moyennes théoriques des populations de référence des deux échantillons observés sont égales,

contre

- H_1 : les moyennes théoriques des populations de référence des deux échantillons observés sont différentes,
- ou H_1' : la moyenne théorique de la population de référence de l'échantillon A est inférieure à la moyenne théorique de la population de référence de l'échantillon B,
- ou H_1'' : la moyenne théorique de la population de référence de l'échantillon A est supérieure à la moyenne théorique de la population de référence de l'échantillon B.

que l'on peut formuler ainsi :

$$\begin{cases} H_0 : \mu_A = \mu_B \\ \text{contre} \begin{cases} H_1 : \mu_A \neq \mu_B \\ \text{ou } H'_1 : \mu_A < \mu_B \\ \text{ou } H''_1 : \mu_A > \mu_B \end{cases} \end{cases}$$

▩ ***Remarque :***

Les hypothèses H'_1 et H''_1 reviennent au même en intervertissant les échantillons A et B.

■ ***Estimation de l'écart réduit calculé ε_c :***

La comparaison entre 2 moyennes m_A et m_B observées sur deux échantillons de taille respective n_A et n_B individus (avec n_A et $n_B \geq 30$) est basée sur l'estimation de l'écart-réduit calculé (c'est-à-dire ε_c) :

$$\varepsilon_c = \frac{m_A - m_B}{\sqrt{\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}}}$$

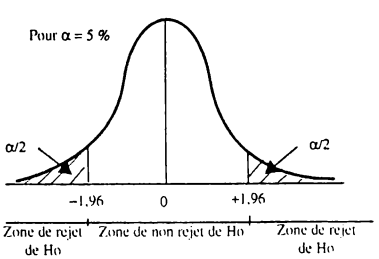
avec s_A^2 et s_B^2 correspondant aux variances estimées de la variable quantitative étudiée, estimées respectivement à partir des échantillons A et B :

$$s_A^2 = \frac{\sum_{i=1}^{n_A} (x_i - m_A)^2}{n_A - 1} \quad \text{et} \quad s_B^2 = \frac{\sum_{i=1}^{n_B} (x_i - m_B)^2}{n_B - 1}$$

1.3.3. Conditions d'application

Le test ε de comparaison de 2 moyennes observées n'est utilisable que si la **taille des échantillons est supérieure ou égale à 30** (avec n_A et $n_B \geq 30$).

➤ ***Dans le cas d'un test bilatéral (H_1)***

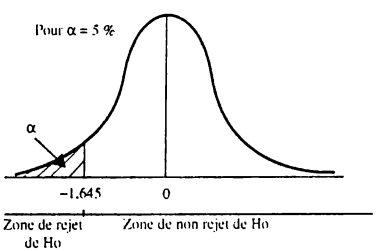


Si $|\varepsilon_c| < 1{,}96$, c'est-à-dire si ε se situe dans la zone de non rejet de H_0, alors on ne rejette pas l'hypothèse H_0. Et on conclut à une différence non significative entre les deux moyennes, au risque $\alpha = 5$ %, avec ε tabulé, ($\varepsilon_t = 1{,}96$), lu dans la table de la loi normale. On conclut alors que les deux échantillons observés sont de moyenne équivalente, c'est-à-dire de moyenne commune μ.

Si $|\varepsilon_c| \geq 1{,}96$, c'est-à-dire si ε se situe dans la zone de rejet de H_0, alors on rejette l'hypothèse H_0. Et on conclut à une différence significative entre les deux moyennes, au risque $\alpha = 5$ %.

Le risque α', lu dans la table de la loi normale, correspondant à la valeur la plus proche de ε_c mais inférieure (en valeur absolue), fixe le degré de signification de la différence (on le note : $p < \alpha'$). Le degré de signification de la différence est le risque α' jusqu'auquel on conclut encore à une différence significative, c'est **le risque minimum de rejeter H_0 à tort**.

➤ *Dans le cas d'un test unilatéral à gauche (H_1')*



Si $\varepsilon_c > -1{,}645$, c'est-à-dire si ε se situe dans la zone de non rejet de H_0, alors on ne rejette pas l'hypothèse H_0. Et on conclut à une inégalité à gauche non significative entre les deux moyennes, au risque $\alpha = 5$ %, avec ε tabulé, ($\varepsilon_t = -1{,}645$), lu dans la table de la loi normale (pour $\alpha_t = 10$ %). On conclut alors que les deux échantillons observés sont de moyenne équivalente, c'est-à-dire de moyenne commune μ.

Si $\varepsilon_c \leq -1{,}645$, c'est-à-dire si ε se situe dans la zone de rejet de H_0, alors on rejette l'hypothèse H_0. Et on conclut à une inégalité à gauche significative entre les deux moyennes, au risque $\alpha = 5$ %. Ainsi, la moyenne A est significativement inférieure à la moyenne B.

Le risque α', lu dans la table de la loi normale, correspondant à la valeur la plus proche de ε_c mais inférieure (en valeur absolue), **fixe le degré de signification de l'inégalité** (on le note : $p < \alpha'$).

➤ *Dans le cas d'un test unilatéral à droite (H''_1)*

On peut se ramener facilement à un test unilatéral à gauche en inversant A et B, puis m_A et m_B dans le calcul de ε_c.

1.4. Comparaison de deux moyennes observées (avec n_A ou $n_B < 30$ et distributions normales de même variance) - test t

1.4.1. Rôle du test t

Le test t permet notamment de comparer **deux échantillons A et B**, de tailles respectives n_A et n_B, à partir de leurs moyennes observées m_A et m_B, dans le cas où n_A ou n_B est **inférieur à 30**. Il s'agit de déterminer si les 2 échantillons observés de moyenne m_A et m_B (concernant la variable quantitative étudiée) peuvent être considérés comme issus de la même population de référence de moyenne théorique commune $\mu = \mu_A = \mu_B$ et de variance théorique commune $\sigma^2 = \sigma_A^2 = \sigma_B^2$.

1.4.2. Procédure du test

Les hypothèses à tester pour comparer deux moyennes observées sont :

H_0 : les moyennes théoriques des populations de référence des deux échantillons observés sont équivalentes,

contre

- H_1 : les moyennes théoriques des populations de référence des deux échantillons observés sont différentes,
- ou H'_1 : la moyenne théorique de la population de référence de l'échantillon A est inférieure à la moyenne théorique de la population de référence de l'échantillon B,
- ou H''_1 : la moyenne théorique de la population de référence de l'échantillon A est supérieure à la moyenne théorique de la population de référence de l'échantillon B,

que l'on peut formuler ainsi :

$$\begin{cases} H_0 : \mu_A = \mu_B \\ \text{contre} \begin{cases} H_1 : \mu_A \neq \mu_B \\ \text{ou } H'_1 : \mu_A < \mu_B \\ \text{ou } H''_1 : \mu_A > \mu_B \end{cases} \end{cases}$$

Remarque :

Les hypothèses H'_1 et H''_1 reviennent au même en intervertissant A et B.

■ ***Calcul de t_c :***

La comparaison entre deux moyennes m_A et m_B observées sur deux échantillons de taille respective n_A et n_B individus (avec n_A ou $n_B < 30$) est basée sur le calcul de t_c :

$$t_c = \frac{m_A - m_B}{\sqrt{\frac{s^2}{n_A} + \frac{s^2}{n_B}}}$$

avec s^2 correspondant à la variance commune estimée à partir des deux échantillons, car les deux variances σ_A^2 et σ_B^2 sont supposées égales :

$$s^2 = \frac{\sum_{i=1}^{n_A}(x_i - m_A)^2 + \sum_{i=1}^{n_B}(x_i - m_B)^2}{n_A + n_B - 2}$$

soit

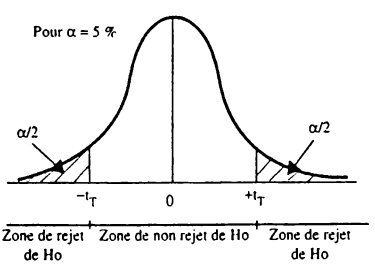
$$s^2 = \frac{s_A^2(n_A - 1) + s_B^2(n_B - 1)}{n_A + n_B - 2}$$

avec $s_A^2 = \frac{\sum_{i=1}^{n_A}(x_i - m_A)^2}{n_A - 1}$ et $s_B^2 = \frac{\sum_{i=1}^{n_B}(x_i - m_B)^2}{n_B - 1}$

1.4.3. Conditions d'application

Le caractère quantitatif étudié doit être distribué, dans les deux populations d'où proviennent les échantillons A et B, selon des **lois normales de même variance.**

➤ ***Dans le cas d'un test bilatéral (H_1)***

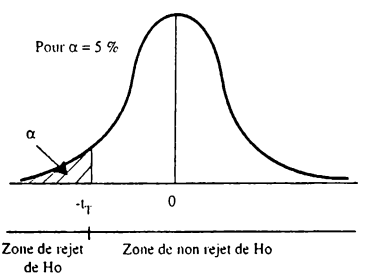


Si $|t_c| < t_T$, c'est-à-dire si t_c se situe dans la zone de non rejet de H_0, alors on ne rejette pas l'hypothèse H_0. Et on conclut à une différence non significative entre les deux moyennes, au risque $\alpha = 5$ %. La valeur de t tabulé, t_T, est lue dans la table de t au risque α choisi pour un degré de liberté égal à $n_A + n_B - 2$.

Si $|t_c| \geq t_T$, c'est-à-dire si t_c se situe dans la zone de rejet de H_0, alors on rejette l'hypothèse H_0. Et on conclut à une différence significative entre les deux moyennes, au risque $\alpha = 5$ %.

Le risque α', lu dans la table de t avec un nombre de degré de liberté égal à $n_A + n_B - 2$, correspondant à la valeur la plus proche de t_c mais inférieure (en valeur absolue), **fixe le degré de signification de la différence** (on le note : $p < \alpha'$). Le degré de signification de la différence est le risque α' jusqu'auquel on conclut encore à une différence significative, c'est **le risque minimum de rejeter H_0 à tort**.

➤ *Dans le cas d'un test unilatéral à gauche (H'_1)*



Si $t_c > -t_T$, c'est-à-dire si t_c se situe dans la zone de non rejet de H_0, alors on ne rejette pas l'hypothèse H_0. Et on conclut à une inégalité à gauche non significative entre les deux moyennes, au risque $\alpha = 5$ %, avec t tabulé, t_T, lu dans la table de t avec un nombre de degré de liberté égal à $n_A + n_B - 2$, pour $\alpha_T = 10$ %.

Si $t_c \leq -t_T$, c'est-à-dire si t_c se situe dans la zone de rejet de H_0, alors on rejette l'hypothèse H_0. Et on conclut à une inégalité à gauche significative entre les deux moyennes, au risque $\alpha = 5$ %, avec t tabulé, t_T, lu dans la table de t avec un nombre de degré de liberté égal à $n_A + n_B - 2$, pour $\alpha_T = 10$ %. Ainsi, la moyenne A est significativement inférieure à la moyenne B.

Le risque α', lu dans la table de t avec ddl = $n_A + n_B - 2$, correspondant à la valeur la plus proche de t_c mais inférieure (en valeur absolue), **fixe le degré de signification de l'inégalité** (on le note : $p < \alpha'$).

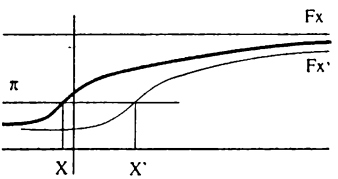
Remarque :

Dans le cas d'un test unilatéral à droite (H''), on peut se ramener facilement à un test unilatéral à gauche en inversant A et B, puis m_A et m_B dans le calcul de t.

1.5. Comparaison de deux lois observées (avec n < 30 et distribution non normale) - test de Mann et Whitney

Soit deux variables aléatoires X et X' respectivement de fonction de répartition F_X et $F_{X'}$, on dit que la variable X est aléatoirement supérieure ou égale à X' si et seulement si $\forall x \in \mathbb{R}\ F_X(X) \geq F_{X'}(X)$.

On le note : $X \geq X' \Leftrightarrow F_X \geq F_{X'}$



1.5.1. Rôle du test de Mann et Whitney

Comparaison de deux caractères observés X et X' à partir de deux échantillons **indépendants**. Les caractères de ces **deux échantillons ne sont pas forcément quantitatifs**, une relation d'ordre suffit. Le test de Mann Whitney permet de comparer les fonctions de répartition de deux échantillons.

1.5.2. Procédure générale du test de Mann et Whitney

Soit deux échantillons, $(x_1, x_2, ..., x_{n1})$ de caractère X de taille n_1, et $(x_1, x_2, ..., x_{n2})$ de caractère X' de taille n_2, les deux échantillons sont indépendants.

Les hypothèses formulées peuvent être :

$$\begin{cases} H_0 : & X \text{ et } X' \text{ sont de même loi} \\ & (\text{i.e. : } F_X = F_{X'} \quad \forall x \in \mathbb{R} F_X(x) = F_{X'}(x)) \\ \text{contre} \begin{cases} & H_1 : X \text{ et } X' \text{ ne sont pas de même loi} \\ & (\text{i.e. : } F_X \neq F_{X'}) \\ \text{ou} & H_1' : X > X' \text{ soit } F_X > F_{X'} \\ \text{ou} & H_1'' : X < X' \text{ soit } F_X < F_{X'} \end{cases} \end{cases}$$

■ *Formation d'un échantillon global*

On rassemble tous les termes des deux échantillons en un échantillon global. On ordonne les termes de cet échantillon global dans l'ordre croissant, en marquant d'un signe distinctif les termes issus du premier échantillon (afin de distinguer les termes des deux échantillons). Puis, dans l'échantillon global, on note le rang de chaque terme.

▩ *Notations :*

On note :

$\mathbf{U_1}$ le nombre de couples (x_i, x'_j) tels que $x'_j < x_i$, c'est-à-dire le nombre de couples dont le terme du deuxième échantillon "précède" le terme du premier échantillon, dans l'échantillon global ;

$\mathbf{U_2}$ le nombre de couples (x_i, x'_j) tels que $x_i < x'_j$, c'est-à-dire le nombre de couples dont le terme du premier échantillon "précède" le terme du deuxième échantillon, dans l'échantillon global ;

$\mathbf{R_1}$ la somme des rangs des termes du premier échantillon ;

$\mathbf{R_2}$ la somme des rangs des termes du deuxième échantillon.

On vérifie le calcul de R_1 et R_2 par la formule de S_n qui correspond à la somme des n premiers rangs, c'est-à-dire

$$S_n = \frac{n(n+1)}{2}$$

et $\quad S_n = R_1 + R_2 \quad$ si $n = n_1 + n_2$

On a les relations suivantes :

$$U_1 = R_1 - \frac{n_1(n_1+1)}{2} \qquad (1)$$

$$U_2 = R_2 - \frac{n_2(n_2+1)}{2} \qquad (2)$$

$$U_1 + U_2 = n_1 n_2 \qquad (3)$$

ou encore

$$U_1 = n_1 n_2 + \frac{n_2(n_2+1)}{2} - R_2 \qquad (4)$$

$$U_2 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1 \qquad (5)$$

à partir des équations (1) et (3) on a alors :

$$U_1 = n_1 n_2 - U_2 = R_1 - \frac{n_1(n_1+1)}{2}$$

où U_1 et U_2 sont des variables aléatoires de même loi.

• *Exemple d'échantillon global :*

Pour l'échantillon global suivant, ordonné de façon croissante

	x'_1	x'_2	x_1	x'_3	x_2	x'_4	x_3
échantillon	II	II	I	II	I	II	I
rang	1	2	3	4	5	6	7

(les indices I et II désignent l'appartenance des termes aux échantillons de départ)

on calcule :

U_1 = nombre de couples (x_i, x'_j) tels que $x'_j < x_i = 2 + 3 + 4 = 9$

U_2 = nombre de couples (x_i, x'_j) tels que $x_i < x'_j = 1 + 2 = 3$

R_1 = somme des rangs des termes du premier échantillon = $3 + 5 + 7 = 15$

R_2 = somme des rangs des termes du deuxième échantillon = $1 + 2 + 4 + 6 = 13$

On peut vérifier les équations (1) et (2) :

$$(1) \Leftrightarrow U_1 = R_1 - \frac{n_1(n_1+1)}{2} \quad \Leftrightarrow \quad 9 = 15 - \frac{3\times 4}{2}$$

$$(2) \Leftrightarrow U_2 = R_2 - \frac{n_2(n_2+1)}{2} \quad \Leftrightarrow \quad 3 = 13 - \frac{4\times 5}{2}$$

■ *Cas de nœuds*

Les définitions de U_1, U_2, R_1 et R_2 correspondent au cas où les termes des échantillons *sont tous distincts.* **Or, il peut arriver que les termes d'un ou des deux échantillons soient égaux : on parle alors de *nœuds*.**

Dans le cas de nœuds, pour assigner le rang de chacun des termes on attribue à chacun des termes égaux de l'échantillon global, le rang correspondant à la moyenne (arithmétique) des rangs naturels de ces termes.

Par exemple, si le 4e et le 5e terme sont égaux, on attribue à chacun d'eux le rang 4,5 soit $\frac{4+5}{2}$. Ou encore, si le 7e, le 8e et le 9e terme sont égaux, on attribue à chacun d'eux le rang $\frac{7+8+9}{3}$ soit 8.

■ *La somme des rangs*

Au lieu de R_1 on note R_1^*, et au lieu de R_2 on note R_2^*. Au lieu de compter le nombre des couples de U_1 et U_2, on compte :

U_1^* = [nombre de couples (x_i, x'_j) tel que $x_i > x'_j$] + $\frac{1}{2}$ [nombre de couples (x_i, x'_j) tel que $x_i = x'_j$]

U_2^* = [nombre de couples (x_i, x'_j) tel que $x_i < x'_j$] + $\frac{1}{2}$ [nombre de couples (x_i, x'_j) tel que $x'_j = x_i$]

On démontre qu'on a des formules analogues :

$$U_1^* = R_1^* - \frac{n_1(n_1+1)}{2} \quad (1')$$

$$U_2^* = R_2 - \frac{n_2(n_2+1)}{2} \quad (2')$$

$$U_1^* + U_2^* = n_1 n_2 \quad (3')$$

• Exemple de somme des rangs pour un échantillon global

	7	8	8	9	10	10	10
échantillon	I	I	II	II	I	II	II
rang	1	2	3	4	5	6	7
rang*	1	2,5	2,5	4	6	6	6

$U_1^* = 2 + \frac{1}{2}(3) = 3,5$

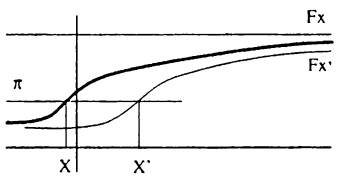
$U_2^* = 1 + 2 + 2 + 2 + \frac{1}{2}(3) = 8,5$

$R_1^* = 1 + 2,5 + 6 = 9,5$
$R_2^* = 2,5 + 4 + 6 + 6 = 18,5$

Selon l'équation (1*) $U_1^* = 3,5 = 9,5 - \frac{3 \times 4}{2} = 3,5$

Selon l'équation (2*) $U_2^* = 8,5 = 18,5 - \frac{4 \times 5}{2} = 8,5$

Remarque :



L'hypothèse X' > X devrait être en correspondance avec le fait que le nombre U_2 de couples (x, x') tel que x < x' *est grand.* En d'autres termes, le nombre U_1 de couples (x, x') tel que x' < x *est petit.*

1.5.3. Test de Mann et Whitney si $n_1 \geq 9$ et $n_2 \geq 9$, échantillons sans nœud

Sous l'hypothèse H_0 : X et X' sont de même loi, la variable aléatoire U_i (avec i = 1, 2) a pour espérance mathématique $E(U_i) = \frac{n_1 n_2}{2}$

et pour variance :

$$\mathrm{Var}(U_i) = \frac{n_1 n_2 (n_1 + n_2 + 1)}{12} \qquad i = 1,2$$

La variable aléatoire V_i :

$$V_i = \frac{U_i - E(U_i)}{\sqrt{V(U_i)}} = \frac{U_i - \frac{n_1 n_2}{2}}{\sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}}}$$

suit une loi normale centrée réduite N(0,1). On utilise donc la table de la loi normale N(0,1), et on note cette variable calculée V_C et la variable lue dans la table de la loi normale V_T.

Les hypothèses étant formulées : il s'agit de calculer U_1 et U_2.
On note $U = \min(U_1, U_2)$
et $\tilde{U} = \max(U_1, U_2)$.

Avec une alternative unilatérale :
- Si on utilise U, calculer V et faire un test unilatéral à gauche.
- Si on utilise $\tilde{U}$, calculer $\tilde{V}$ et faire un test unilatéral à droite.

Avec une alternative bilatérale :
- Si on utilise U, calculer V et faire une lecture unilatérale à gauche.
- Si on utilise $\tilde{U}$, calculer $\tilde{V}$ et faire une lecture unilatérale à droite.

• Exemple :

Le tableau suivant donne le score de qualité de vie de deux échantillons de sujets ayant répondu à une échelle de qualité de vie.

Echantillon I	14	13,2	18,4	14,5	15,1	23,6	14,7	12
X		10,2	14,3	11,8	14,9	13,5	10,7	16,3
Echantillon II	16,2	15,8	21,1	20,7	18,9	14,8	20,1	15,3
X'		19,2	18,8	13,9	12,6	21,3	19,6	15,4

Tester les hypothèses suivantes au seuil $\alpha = 1\ \%$:

H_0 : les deux échantillons sont équivalents en terme de qualité de vie (X et X' sont de même loi)

H_1 : les sujets de l'échantillon II obtiennent de meilleurs scores de qualité de vie que ceux de l'échantillon I (loi X < loi de X')

L'échantillon global étant sans nœud, on calcule R_1 et R_2, par les formules (1) et (2), et on en déduit U_1 et U_2.
On forme un échantillonnage global en calculant les rangs, puis U_1 et U_2.

Echantillon global ordonné :

	10,2	10,7	11,8	12	12,6	13,2	13,5	13,9	14	14,3	14,5
Rang	1	2	3	4	5	6	7	8	9	10	11
Echantillon	I	I	I	I	II	I	I	II	I	I	I

	14,7	14,8	14,9	15,1	15,3	15,4	15,8	16,2	16,3	18,4	18,8
Rang	12	13	14	15	16	17	18	19	20	21	22
Echantillon	I	II	I	I	II	II	II	II	I	I	II

	18,9	19,2	19,6	20,1	20,7	21,1	21,3	23,6
Rang	23	24	25	26	27	28	29	30
Echantillon	II	II	II	II	II	II	II	I

R_1 = somme des rangs de l'échantillon I = 165
R_2 = somme des rangs de l'échantillon II = 300

On vérifie que $S_n = R_1 + R_2$

soit $S_n = \frac{n(n+1)}{2} = \frac{30 \times 31}{2} = 15 \times 31 = 465$

et on en déduit :

$$U_1 = R_1 - \frac{n_1(n_1+1)}{2} = 165 - \frac{15 \times 16}{2} = 45$$

$$U_2 = R_2 - \frac{n_2(n_2+1)}{2} = 300 - \frac{15 \times 16}{2} = 180$$

on a bien $U_1 + U_2 = n_1 n_2 \Leftrightarrow 45 + 180 = 15 \times 15 = 225$

Alors $U = \min(U_1, U_2) = 45$
et $\tilde{U} = \max(U_1, U_2) = 180$

Les hypothèses à tester étant :

$$\begin{cases} H_0 : X = X' \\ H_1 : X < X' \end{cases}$$

ou encore

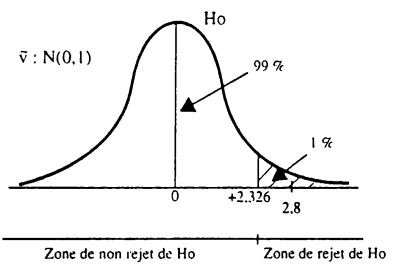
$$\begin{cases} H_0 : X' = X \\ H_1 : X' > X \end{cases}$$

Le nombre U_2 de couples (x, x') tel que x < x' très grand, équivalent à 180, est compatible avec l'hypothèse alternative H_1 : X' > X qui correspond à un test unilatéral à droite.

Alors $\tilde{U} = U_2 = 180$

$$\text{Donc} \quad \tilde{V}_C = \frac{\tilde{U} - \frac{n_1 n_2}{2}}{\sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}}} = \frac{180 - 112{,}5}{\sqrt{\frac{15 \times 15 \times 31}{12}}} = 2{,}8$$

La variable tabulée $\tilde{V}_1$ est lue dans la table de la loi normale au seuil $\alpha = 1$ % (la table présentée en annexe correspondant au cas bilatéral, on lit $\tilde{V}_T$ pour $\alpha = 2$ %, soit $\tilde{V}_T = 2{,}326$).



La variable calculée $\tilde{V}_C$ se situant dans la zone de rejet de H_0 (on a $\tilde{V}_T < \tilde{V}_C$), on rejette l'hypothèse H_0 au seuil $\alpha = 1$ %. Ainsi, on conclut que l'échantillon II obtient un meilleur score de qualité de vie que l'échantillon I (X' > X) et le degré de signification de cette inégalité est $p < 0{,}005$ (car jusqu'au seuil $\alpha = 0{,}5$ % on rejette encore l'hypothèse nulle).

1.5.4. Test de Mann et Whitney si $n_1 \geq 9$ et $n_2 \geq 9$, échantillons avec nœud

Dans ce cas on peut toujours utiliser la loi Normale pour la variable calculée. La variable aléatoire U_i^* a pour espérance mathématique :

$$E(U_i^*) = \frac{n_1 n_2}{2}$$

et pour variance : $$\text{Var}(U_i^*) = \frac{n_1 n_2 (N+1)}{12} - \frac{n_1 n_2 \sum_{i=1}^{e} (d_i^3 - d_i)}{12 N(N-1)}$$

avec e : le nombre de termes différents de l'échantillon global ;
N : la taille de l'échantillon global, soit $N = n_1 + n_2$;
d_i : le nombre de terme ayant la i^e valeur.

Par exemple pour les termes	22	4	999
on a	$d_1 = 2$	$d_2 = 1$	$d_3 = 3$

Ainsi on utilise la formule :

$$V^* = \frac{U_i^* - \frac{n_1 n_2}{2}}{\sqrt{Var(U_i^*)}} = \frac{U_i^* - E(U_i^*)}{\sqrt{Var(U_i^*)}}$$

et V^* suit une loi Normale N(0, 1). On note cette variable calculée V_c^*, et la variable tabulée, notée V_T^*, est lue dans la table de la loi normale au seuil α choisi (le plus souvent $\alpha = 5\ \%$).

Si $V_c^* > V_T^*$ alors on rejette l'hypothèse nulle H_0.

• Exemple :

On recueille les temps de réponse (en seconde) à une question chez deux groupes de sujets de taille respective $n_1 = 16$ et $n_2 = 23$. Les données sont les suivantes :

Echantillon 1 : X	13	12	12	10	10	10	10	9	8	8
	7	7	7	7	7	6				
Echantillon 2 : X'	17	16	15	15	15	14	14	14	13	13
	13	12	12	12	12	11	11	10	10	10
	8	8	6							

Tester au seuil $\alpha = 1\ \%$, les hypothèses suivantes :

H_0 : X = X' la loi de l'échantillon 1 est équivalente à celle de l'échantillon 2.
H_1 : X < X' la loi de l'échantillon 1 est stochastiquement inférieure à celle de l'échantillon 2.

Echantillon global ordonné :

	6	6	7	7	7	7	7	8	8	8	8
Echantillon	1	2	1	1	1	1	1	1	1	2	2
Rang naturel	1	2	3	4	5	6	7	8	9	10	11
Rang	1,5	1,5	5	5	5	5	5	9,5	9,5	9,5	9,5

	9	10	10	10	10	10	10	10	11	11
Echantillon	1	1	1	1	1	2	2	2	2	2
Rang naturel	12	13	14	15	16	17	18	19	20	21
Rang	12	16	16	16	16	16	16	16	20,5	20,5

	12	12	12	12	12	12	13	13	13	13
Echantillon	1	1	2	2	2	2	1	2	2	2
Rang naturel	22	23	24	25	26	27	28	29	30	31
Rang	24,5	24,5	24,5	24,5	24,5	24,5	29,5	29,5	29,5	29,5

	14	14	14	15	15	15	16	17
Echantillon	2	2	2	2	2	2	2	2
Rang naturel	32	33	34	35	36	37	38	39
Rang	33	33	33	36	36	36	38	39

■ *Calcul des rangs :*

Pour les termes 6 6, on compte $\frac{1+2}{2} = 1,5$ $d_1 = 2$

pour 7 7 7 7 7, on compte $\frac{3+4+5+6+7}{5} = 5$ $d_2 = 5$

pour 8 8 8 8, on compte $\frac{8+9+10+11}{4} = 9,5$ $d_3 = 4$

pour 9, le rang est 12 $d_4 = 1$

pour 10 10 10 10 10 10 10, on compte

$\frac{13+14+15+16+17+18+19}{7} = 16$ $d_5 = 7$

pour 11 11, on compte $\frac{20+21}{2} = 20,5$ $d_6 = 2$

pour 12 12 12 12 12 12, on compte

$$\frac{22+23+24+25+26+27}{6} = 24{,}5 \qquad d_7 = 6$$

et ainsi de suite

$d_8 = 4$
$d_9 = 3$
$d_{10} = 3$
$d_{11} = 1$
$d_{12} = 1$

■ ***Calcul de R_1^* et R_2^* :***

$$R_1^* = 1{,}5 + (5 \times 5) + (9{,}5 \times 2) + 12 + (16 \times 4) + (24{,}5 \times 2) + 29{,}5 = 200$$

$$R_2^* = 1{,}5 + (9{,}5 \times 2) + (16 \times 3) + (20{,}5 \times 2) + (24{,}5 \times 4) + (29{,}5 \times 3) + (33 \times 3) + (36 \times 3) + 38 + 39 = 580$$

On a :

$$R_1^* + R_2^* = 780$$

$$= n_1 n_2 + \frac{n_1(n_2+1)}{2} + \frac{n_2(n_1+1)}{2}$$

$$= n_1\left[n_2 + \frac{n_2+1}{2} + n_2\right] + \frac{n_2(n_1+1)}{2}$$

$$= \frac{2n_1n_2 + n_1n_2 + n_1 + n_1n_2 + n_2}{2}$$

$$= \frac{4n_1n_2 + n_1 + n_2}{2}$$

Ainsi $R_1^* + R_2^* = \frac{4n_1n_2 + n_1 + n_2}{2}$ permet de vérifier les calculs.

■ ***Calcul de U_1^* et U_2^* :***

$$U_1^* = R_1^* - \frac{n_1(n_1+1)}{2} = 200 - \frac{16 \times 17}{2} = 64$$

$$U_2^* = R_2^* - \frac{n_2(n_2+1)}{2} = 580 - \frac{23 \times 24}{2} = 304$$

En considérant U = 64, on réalise alors un test unilatéral à gauche qui est en correspondance avec l'hypothèse alternative H_1.

Donc on calcule :

$$V^* = \frac{U_i^* - E(U_i^*)}{\sqrt{Var(U_i^*)}} = \frac{U_i^* - \frac{n_1 n_2}{2}}{\sqrt{\frac{n_1 n_2 (N+1)}{12} - \frac{n_1 n_2 \sum_{i=1}^{e} (d_i^3 - d_i)}{12 N(N-1)}}}$$

Pour cela on calcule les $d_i^3 - d_i$:

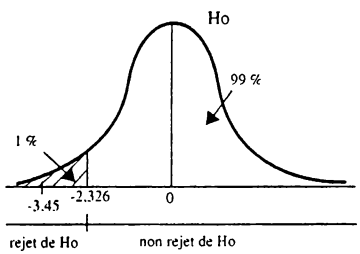
$i = 1 \rightarrow 2^3 - 2$	$i = 4 \rightarrow 1^3 - 1$	$i = 7 \rightarrow 6^3 - 6$	$i = 10 \rightarrow 3^3 - 3$
$i = 2 \rightarrow 5^3 - 5$	$i = 5 \rightarrow 7^3 - 7$	$i = 8 \rightarrow 4^3 - 4$	$i = 11 \rightarrow 1^3 - 1$
$i = 3 \rightarrow 4^3 - 4$	$i = 6 \rightarrow 2^3 - 2$	$i = 9 \rightarrow 3^3 - 3$	$i = 12 \rightarrow 1^3 - 1$

Alors $\sum_{i=1}^{e} (d_i^3 - d_i) = 846$

Ainsi

$$V^* = \frac{64 - \frac{16 \times 23}{2}}{\sqrt{\frac{16 \times 23 (16+23+1)}{12} - \frac{16 \times 23 \times 846}{12(16+23)(16+23-1)}}} = \frac{-120}{\sqrt{1226,66 - 17,506}} = -3,45$$

Au seuil $\alpha = 1$ % le paramètre tabulé, lu dans la table de la loi Normale, est $V_T = -2,33$



La valeur du paramètre calculé étant inférieure à celle du paramètre tabulé (c'est-à-dire se situant dans la zone de rejet de H_0), alors on rejette l'hypothèse H_0. Et on conclut que la loi du temps de réponse du premier échantillon est aléatoirement inférieure à celle du deuxième échantillon. C'est-à-dire que les sujets observés du premier groupe (de taille $n_1 = 16$) répondent plus rapidement à la question étudiée, que ceux du second groupe (de taille $n_2 = 23$).

1.5.5. Test de Mann et Whitney si $n_1 \leq 8$ et $n_2 \leq 8$, et échantillons sans nœud

La procédure du test est la même que dans le 1er cas en utilisant ici la table de Mann et Whitney (jointe en annexe). On se place toujours dans le cas où $n_1 \leq n_2$.

Par exemple si $n_1 = 2$ et $n_2 = 3$, alors $n_1 n_2 = 6$, puis U_1 et U_2 peuvent prendre les valeurs 0, 1, 2, 3, 4, 5 ou 6, car $U_1 + U_2 = n_1 n_2$.

Si $U_1 = 0$ alors $U_2 = 6$
$U_1 = 1$ alors $U_2 = 5$
$U_1 = 2$ alors $U_2 = 4$
$U_1 = 3$ alors $U_2 = 3$
$U_1 = 4$ alors $U_2 = 2$
$U_1 = 5$ alors $U_2 = 1$
$U_1 = 6$ alors $U_2 = 0$

On peut lire dans la table de Mann et Whitney, au seuil $\alpha = 5$ % (ou au seuil $\alpha = 1$ %), la valeur de U_T à l'intersection de la colonne n_1 et de la ligne $n_2 - n_1$. Puis si la valeur de U calculée, U_C, est inférieure ou égale à la valeur de U tabulé, U_T, alors on conclut au rejet de H_0, c'est-à-dire à une différence significative.

Donc selon cet exemple, U_C peut prendre les valeurs 0, 1, 2 ou 3. Or, de toutes façons selon la table pour les effectifs $n_1 = 2$ et $n_2 = 3$, la différence n'est jamais significative au seuil $\alpha = 5$ %.

• Exemple :

Soit deux échantillons de taille respective $n_1 = 2$ et $n_2 = 4$. Les hypothèses à tester, au seuil $\alpha = 5$ %, sont : H_0 : X et X' sont de même loi
contre H_1 : X et X' sont de loi différente
On a $U_1 = 2$ et $U_2 = 6$, calculés par les méthodes vues précédemment.
La table de Mann et Whitney ne donne pas la valeur de U_T car pour des effectifs $n_1 = 2$ et $n_2 = 4$ la différence n'est jamais significative. Donc on rejette l'hypothèse H_1 et on conclut que les variables X et X' sont de même loi.

1.5.6. Test de Mann et Whitney si $n_1 \leq 8$ et $9 \leq n_2 \leq 20$ et échantillon sans nœud

La table de Mann et Whitney donne la valeur U_T correspondant à n_1 et $n_2 - n_1$ que l'on compare à la valeur calculée U_C. Si $U_C \leq U_T$, au seuil $\alpha = 5$ %, alors on rejette l'hypothèse nulle et on conclut à une différence significative.

• Exemple 1 :

Pour $n_1 = 6$ et $n_2 = 13$ au seuil $\alpha = 0{,}05 = 5\ \%$, on lit U_T à l'intersection de la colonne 6 et de la ligne $n_2 - n_1 = 7$, soit $U_T = 16$.
Si $U_C = 8$ alors $U_C \leq U_T$ donc on rejette l'hypothèse H_0.
Pour $n_1 = 6$ et $n_2 = 13$ au seuil $\alpha = 0{,}01 = 1\ \%$, on lit U_T à l'intersection de la colonne 6 et de la ligne $n_2 - n_1 = 7$, soit $U_T = 10$.
Si $U_C = 8$ alors $U_C \leq U_T$ donc on rejette encore l'hypothèse H_0, donc $p < 0{,}01$.

• Exemple 2 :

Dans le cas où les hypothèses à tester sont

$$\begin{cases} H_0 : \text{X et X' sont de même loi} \\ \text{contre } H_1 : \text{X et X' sont de loi différente} \end{cases}$$

Si $n_1 = 7$ et $n_2 = 10$ au seuil $\alpha = 0{,}05 = 5\ \%$, on a $U_T = 14$.
Si $U_C > 14$, c'est-à-dire si $U_C > U_T$ alors on rejette l'hypothèse H_1, et on conclut à une différence non significative.

Remarque :

La table de Mann et Whitney doit être lue de façon à avoir toujours $n_1 \leq n_2$.

1.5.7. Test de Mann et Whitney si $n_1 \leq 8$ et $n_2 > 20$, et échantillons sans nœud ou avec nœud

Dans ce cas on utilise toujours la table de la loi normale, et on procède comme dans le premier ou le deuxième cas selon que l'on ne dispose pas ou que l'on dispose de nœuds.

1.5.8. Test de Mann et Whitney si $n_1 \leq 8$ et $n_2 \leq 8$, et échantillons avec nœud

Selon Gibons, lorsque les nœuds sont à l'intérieur d'un échantillon ou à l'intérieur de chacun des deux échantillons (nœuds internes), on fait comme si les nœuds n'existaient pas, car dans ce cas les nœuds n'interviennent pas dans le calcul de U_1 et U_2, et on utilise la table de Mann et Whitney.

Lorsque les nœuds sont des nœuds croisés (obtenus seulement dans l'échantillon global), alors on élimine les termes formant des nœuds. Cependant cette méthode n'est pas satisfaisante car il ne reste que peu de termes.

• Exemple :

Dans le cas où les hypothèses à tester, au seuil $\alpha = 5\ \%$, sont

$$\begin{cases} H_0 : \text{X et X' sont de même loi} \\ \text{contre } H_1 : \text{X et X' sont de loi différente} \end{cases}$$

pour deux échantillons indépendants de taille $n_1 = 4$ et $n_2 = 4$,

soit :					
	échantillon 1 (X)	1	3	4	10
	échantillon 2 (X')	6	6	6	9

L'échantillon global ordonné est alors :

	1	3	4	6	6	6	9	10
échantillon	1	1	1	2	2	2	2	1

Ainsi $U_1 = 4 = U_1^*$ et $U_2 = 12 = U_2^*$

On considère la plus petite des deux valeurs U_1 et U_2, soit ici $U_C = U_1 = 4$, que l'on compare à U_T au seuil $\alpha = 5\ \%$. On lit U_T dans la table de Mann et Whitney à l'intersection de la colonne 4 et de la ligne 0, soit $U_T = 0$. Donc, $U_C > U_T$, alors on rejette l'hypothèse H_1 et on conclut à une différence non significative.

1.6. Comparaison de deux variances

1.6.1. Rôle du test de comparaison de deux variances

Le test de comparaison de deux variances :

1) Permet de vérifier les conditions d'application du test t lors de la comparaison de deux moyennes, lorsque les échantillons sont petits (au moins un des deux échantillons est de taille $n < 30$) afin de vérifier l'égalité des variances ;

2) Permet de vérifier les conditions d'application de l'analyse de variance lors de la comparaison de plusieurs moyennes, également afin de vérifier l'égalité des variances deux à deux ;

3) Permet de comparer la précision de deux méthodes, par la comparaison de leur variance, quand une méthode est jugée d'autant plus précise que la variabilité est plus faible.

Il s'agit, lors de la comparaison des variances, de deux échantillons, de déterminer si les variances théoriques du caractère quantitatif étudié des deux populations de référence dont sont issus ces échantillons sont égales ou non.

1.6.2. Procédure du test

Les hypothèses à tester sont :

H_0 : les variances théoriques des populations de référence des deux échantillons observés sont équivalentes,

contre
- H_1 : les variances théoriques des populations de référence des deux échantillons observés sont différentes,
- ou H'_1 : la variance théorique de la population de référence de l'échantillon A est inférieure à la variance théorique de la population de référence de l'échantillon B,
- ou H''_1 : la variance théorique de la population de référence de l'échantillon A est supérieure à la variance théorique de la population de référence de l'échantillon B,

que l'on peut formuler ainsi :

$H_0 : \sigma_A^2 = \sigma_B^2$

contre
- $H_1 : \sigma_A^2 \neq \sigma_B^2$
- ou $H'_1 : \sigma_A^2 < \sigma_B^2$
- ou $H''_1 : \sigma_A^2 > \sigma_B^2$

■ ***Remarque :***

Les hypothèses H'_1 et H''_1 reviennent au même en inversant les échantillons A et B.

■ ***La comparaison entre deux variances :***

La comparaison entre 2 variances s_A^2 et s_B^2 observées sur deux échantillons de tailles respectives n_A et n_B individus, extraits de deux populations pour lesquelles le caractère étudié est distribué normalement, est basée sur l'estimation du rapport F :

$$F_c = \frac{s_A^2}{s_B^2}$$

En effet, si l'hypothèse H_0 est vérifiée, c'est-à-dire si les variances théoriques σ_A^2 et σ_B^2 sont équivalentes, alors le rapport F_c (F calculé) doit être proche de 1 et rester compris entre des limites inférieure et supérieure F_i et F_s (autour de 1), avec un risque d'erreur $\alpha = 5\ \%$ (risque de rejet de H_0 à tort). Donc si F_c est en dehors de cet intervalle $]F_i ; F_s[$, on conclut que les variances diffèrent significativement. Plus simplement, en prenant

toujours la plus grande des deux variances estimées au numérateur, il suffit de comparer F_c à F_s. On note alors F_s plutôt que F_T, pour F tabulé (lu dans la table de F).

1.6.3. Lecture des tables de F

Les tables de F jointes en annexe résument la loi de probabilité de F et permettent de déterminer F_s (ou F_T) pour différents risques d'erreur de première espèce (α), en fonction des degrés de liberté (n_A - 1) lu en colonne et (n_B - 1) lu en ligne dans les tables.

Dans le cas d'un test bilatéral (H_1), pour un risque α choisi à 5 %, la lecture de F_T doit être effectuée dans la table "point 2,5 %", à l'intersection de la colonne (n_A - 1) et de la ligne (n_B - 1), car les tables de F correspondent au cas unilatéral.

Dans le cas unilatéral à droite (H_1''), pour un risque α choisi à 5 %, la lecture de F_T doit être effectuée dans la table "point 5 %" à l'intersection de la colonne (n_A - 1) et de la ligne (n_B - 1). Ce cas sera privilégié par rapport au cas unilatéral à gauche, puisque l'on considère toujours la plus grande des deux variances estimée au numérateur notée s_A^2.

1.6.4. Procédure résumée

La comparaison de deux variances s_A^2 et s_B^2 estimées sur n_A et n_B individus (s_A^2 désignant la plus grande de ces deux variances) est basée sur le calcul du rapport :

$$F_c = \frac{s_A^2}{s_B^2}$$

que l'on compare à la valeur F_T lue dans la table de F à l'intersection de la colonne (n_A - 1) et de la ligne (n_B - 1).

Si $F_c < F_T$, on rejette l'hypothèse alternative, les deux variances ne diffèrent pas significativement au risque 5 %.

Si $F_c \geq F_T$, on rejette l'hypothèse nulle H_0, les deux variances diffèrent significativement (dans le cas d'un test bilatéral), ou la variance A est significativement supérieure à la variance B (dans le cas d'un test unilatéral à droite). Puis le degré de signification de cette différence ou de cette inégalité peut être précisé en comparant F_c à la valeur de F_T lue dans les tables "point 2,5 %" (signification à 2,5 % dans le cas unilatéral), "point 1 %" (signification à 1% dans le cas unilatéral, signification à 2 % dans le cas bilatéral) et "point 1 ‰" (signification à 1 ‰ dans le cas unilatéral, signification à 2 ‰ dans le cas bilatéral).

1.6.5. Conditions d'application

Le test de comparaison de deux variances n'est applicable que si les deux échantillons sont issus de populations pour lesquelles le caractère étudié est distribué normalement.

2. Variables quantitatives dans le cas de deux groupes avec données appariées

2.1. Comparaison de deux moyennes observées (avec n ≥ 30 et distribution normale) - test ε (ou test z)

2.1.1. Rôle du test

Le test ε permet de comparer **deux moyennes observées** m_A et m_B, dans le cas de **deux séries de données appariées de grande taille**, n supérieure ou égale à 30. Il s'agit de déterminer si les deux échantillons appariés observés de moyenne m_A et m_B (concernant la variable quantitative étudiée) peuvent être considérés ou non comme issus de la même population de référence de moyenne théorique commune $\mu = \mu_A = \mu_B$.

2.1.2. Procédure du test

Les hypothèses à tester sont :

H_0 : les moyennes théoriques des populations de référence des deux échantillons observés sont équivalentes,

contre
- H_1 : les moyennes théoriques des populations de référence des deux échantillons observés sont différentes,
- H'_1 : la moyenne théorique de la population de référence de l'échantillon A est inférieure à la moyenne théorique de la population de référence de l'échantillon B,
- H''_1 : la moyenne théorique de la population de référence de l'échantillon A est supérieure à la moyenne théorique de la population de référence de l'échantillon B,

que l'on peut formuler ainsi :

$$\begin{cases} H_0 : \mu_A = \mu_B \\ \text{contre} \begin{cases} H_1 : \mu_A \neq \mu_B \\ \text{ou } H'_1 : \mu_A < \mu_B \\ \text{ou } H''_1 : \mu_A > \mu_B \end{cases} \end{cases}$$

Cela revient également à étudier la moyenne des différences entre les mesures, m_d. Plus cette moyenne est proche de 0, moins les donnés appariées diffèrent dans l'ensemble. C'est pourquoi, on compare la moyenne des différences, m_d, à 0. Les hypothèses à tester sont alors :

$$\begin{cases} H_0 : \mu_d = 0 \\ \text{contre} \begin{cases} H_1 : \mu_d \neq 0 \\ \text{ou } H'_1 : \mu_d < 0 \\ \text{ou } H''_1 : \mu_d > 0 \end{cases} \end{cases}$$

avec μ_d : la moyenne théorique des différences

Pour comparer les moyennes observées de deux séries appariées, on calcule pour chaque couple de données observées la différence des mesures afin de déterminer la moyenne des différences, m_d.

Quand on compare un échantillon à lui-même (concernant la variable quantitative étudiée), à deux temps différents, ou avec deux types de mesure différents, ou deux méthodes différentes, on dispose de deux séries de données appariées pour lesquelles on peut calculer la différence de chaque couple d'observations, et par suite la moyenne de ces différences.

• Exemple :

Temps de réponse de n sujets à deux moments différents, t_A et t_B

t_A	t_B	$d = t_A - t_B$
t_{1A}	t_{1B}	$d_1 = t_{1A} - t_{1B}$
t_{2A}	t_{2B}	$d_2 = t_{2A} - t_{2B}$
t_{3A}	t_{3B}	$d_3 = t_{3A} - t_{3B}$
...	...	
t_{iA}	t_{iB}	$d_i = t_{iA} - t_{iB}$
...	...	
t_{nA}	t_{nB}	$d_n = t_{nA} - t_{nB}$
		Σd_i

■ ***Calcul de la moyenne des différences :***

Pour chacun des n couples, on calcule la différence entre deux mesures et on obtient $d_1, d_2, ..., d_n$, puis on calcule la moyenne des différences par la formule :

$$m_d = \frac{\sum_{i=1}^{n} d_i}{n}$$

soit $$m_d = \frac{\text{Somme des différences}}{\text{Nombre de sujets}}$$

■ *Calcul de l'écart-réduit :*

Au lieu de comparer deux moyennes, on compare une moyenne des différences m_d à une moyenne des différences nulle. Ainsi la comparaison entre une moyenne et 0 est basée sur le calcul de l'écart-réduit (c'est-à-dire ε_c) :

$$\varepsilon_c = \frac{m_d - 0}{\frac{s_d}{\sqrt{n}}}$$

soit

$$\varepsilon_c = \frac{m_d}{\frac{s_d}{\sqrt{n}}}$$

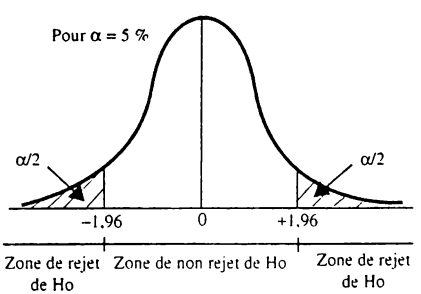
avec s_d correspondant à l'écart-type de la variable quantitative étudiée, estimé à partir de l'échantillon des n différences :

$$s_d = \sqrt{\frac{\sum_{i=1}^{n} (d_i - m_d)^2}{n-1}}$$

2.1.3. Conditions d'application

La taille des séries appariées, **n, doit être supérieure ou égale à 30** ($n \geq 30$).

➤ *Dans le cas d'un test bilatéral (H_1) :*

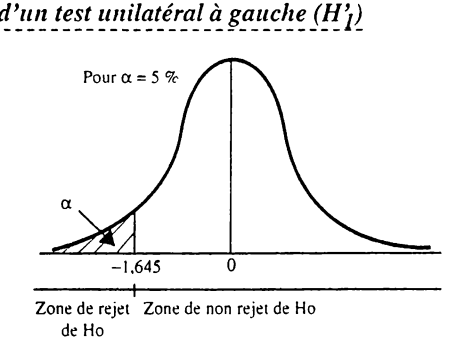


Si $|\varepsilon_c| < 1{,}96$ c'est-à-dire si ε se situe dans la zone de non rejet de H_0, alors on ne rejette pas l'hypothèse H_0. Et on conclut à une différence non significative entre les deux moyennes, μ_A et μ_B, au risque $\alpha = 5\ \%$, avec ε tabulé, ($\varepsilon_t = 1{,}96$), lu dans la table de la loi normale. On conclut alors que les deux séries appariées sont de moyennes équivalentes, c'est-à-dire de moyenne commune μ.

Si $|\varepsilon_c| \geq 1{,}96$ c'est-à-dire si ε se situe dans la zone de rejet de H_0, alors on rejette l'hypothèse H_0 et on conclut à une différence significative entre les moyennes au risque $\alpha = 5\ \%$.

Le risque α', lu dans la table de la loi normale, correspondant à la valeur la plus proche de ε_c mais inférieure (en valeur absolue), fixe le degré de signification de la différence (on le note : $p < \alpha'$). Le degré de signification de la différence est le risque α' jusqu'auquel on conclut encore à une différence significative, c'est le **risque minimum de rejeter H_0 à tort**.

➤ *Dans le cas d'un test unilatéral à gauche (H'_1)*



Si $\varepsilon_c > -1{,}645$, c'est-à-dire si ε se situe dans la zone de non rejet de H0, alors on ne rejette pas l'hypothèse H_0. Et on conclut à une inégalité à gauche non significative entre les deux moyennes, au risque $\alpha = 5\ \%$, avec ε tabulé, ($\varepsilon_t = -1{,}645$), lu dans la table de la loi normale (pour $\alpha_t = 10\ \%$). On conclut alors que les deux séries appariées sont de moyennes équivalentes, c'est-à-dire de moyenne commune μ.

Si $\varepsilon_c \leq -1{,}645$, c'est-à-dire si ε se situe dans la zone de rejet de H_0, alors on rejette l'hypothèse H_0 et on conclut à une inégalité à gauche significative entre les deux moyennes, au risque $\alpha = 5\ \%$. Ainsi, la moyenne A est significativement inférieure à la moyenne B.

Le risque α', lu dans la table de la loi normale, correspondant à la valeur la plus proche de ε_c mais inférieure (en valeur absolue) **fixe le degré de signification de l'inégalité** (on le note : $p < \alpha'$).

➤ *Dans le cas d'un test unilatéral à droite (H_1'')*

On peut se ramener facilement à un test unilatéral à gauche en inversant A et B, dans le calcul de m_d et s_d.

Sinon, si $\varepsilon_c < +1{,}645$ on ne rejette pas l'hypothèse H_0, au risque $\alpha = 5\ \%$.

Puis, si $\varepsilon_c \geq +1{,}645$ on rejette l'hypothèse H_0, au risque $\alpha = 5\ \%$. Et on conclut que la moyenne A est significativement supérieure à la moyenne B. Puis de la même façon, le risque α' fixe le degré de signification de l'inégalité.

2.2. Comparaison de deux moyennes observées (avec n < 30 et distribution normale) - test t

2.2.1. Rôle du test

Le test t permet de comparer **deux moyennes observées** m_A et m_B, dans le cas de deux séries de **données appariées de petite taille**, n inférieure à 30. Il s'agit de déterminer si les deux échantillons appariés observés de moyenne m_A et m_B (concernant la variable quantitative étudiée) peuvent être considérés ou non comme issus de la même population de référence de moyenne théorique commune $\mu = \mu_A = \mu_B$.

2.2.2. Procédure du test

Les hypothèses à tester sont :

H_0 : les moyennes théoriques des populations de référence des deux échantillons observés sont équivalentes,

contre
- H_1 : les moyennes théoriques des populations de référence des deux échantillons observés sont différentes,
- H_1' : la moyenne théorique de la population de référence de l'échantillon A est inférieure à la moyenne théorique de la population de référence de l'échantillon B,
- H_1'' : la moyenne théorique de la population de référence de l'échantillon A est supérieure à la moyenne théorique de la population de référence de l'échantillon B,

que l'on peut formuler ainsi :

$H_0 : \mu_A = \mu_B$

contre
- H_1 : $\mu_A \neq \mu_B$
- ou H_1' : $\mu_A < \mu_B$
- ou H_1'' : $\mu_A > \mu_B$

Cela revient également à étudier la moyenne des différences entre les mesures, m_d. Plus cette moyenne est proche de 0, moins les données appariées diffèrent dans l'ensemble. C'est pourquoi on compare la moyenne des différences, m_d, à 0. Les hypothèses à tester sont alors :

$$\begin{cases} H_0 : \mu_d = 0 \\ \text{contre} \begin{cases} H_1 : \mu_d \neq 0 \\ \text{ou } H_1' : \mu_d < 0 \\ \text{ou } H_1'' : \mu_d > 0 \end{cases} \end{cases} \quad \text{avec } \mu_d \text{ : la moyenne théorique des différences}$$

Pour comparer les moyennes observées de deux séries appariées, on calcule pour chaque couple de données observées la différence des mesures afin de déterminer la moyenne des différences, m_d (voir calcul de m_d dans l'exemple du paragraphe 2.1.2.).

■ ***Calcul de t_c :***

Au lieu de comparer deux moyennes, on compare une moyenne des différences à 0. Ainsi la comparaison entre une moyenne et 0 est basée sur le calcul de t (t_c pour t calculé) :

$$t_c = \frac{m_d - 0}{\frac{s_d}{\sqrt{n}}}$$

soit

$$\boxed{t_c = \frac{m_d}{\frac{s_d}{\sqrt{n}}}}$$

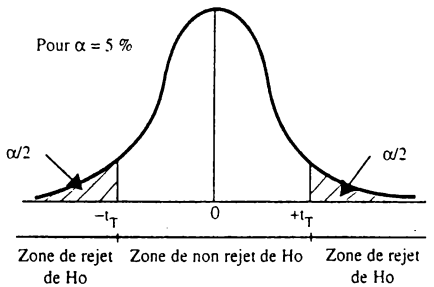
avec s_d correspondant à l'écart-type de la variable quantitative étudiée, estimé à partir de l'échantillon des n différences :

$$\boxed{s_d = \sqrt{\frac{\sum_{i=1}^{n} (d_i - m_d)^2}{n-1}}}$$

2.2.3. Conditions d'application

La différence entre les mesures doit être **distribuée normalement**.

➤ *Dans le cas d'un test bilatéral (H_1)*

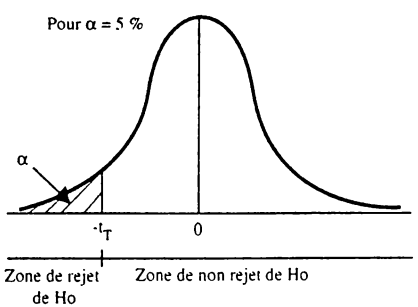


Si $|t_c| < t_T$, c'est-à-dire si t_c se situe dans la zone de non rejet de H_0, alors ne rejette pas l'hypothèse H_0. Et on conclut à une différence non significative entre les deux moyennes, μ_A et μ_B, au risque $\alpha = 5$ %, avec t_T, lu dans la table de t, avec un nombre de degré de liberté (n - 1). On conclut alors que les deux séries appariées sont de moyennes équivalentes, c'est-à-dire de moyenne commune μ.

Si $|t_c| \geq t_T$, c'est-à-dire si t_c se situe dans la zone de rejet de H_0, alors on rejette l'hypothèse H_0 et on conclut à une différence significative entre les moyennes, au risque $\alpha = 5$ %.

Le risque α', lu dans la table de t, avec un nombre de degré de liberté égal à (n - 1), correspondant à la valeur la plus proche de t_c mais inférieure (en valeur absolue), fixe le degré de signification de la différence (on le note : $p < \alpha'$). Le degré de signification de la différence est le risque α' jusqu'auquel on conclut encore à une différence significative, **c'est le risque minimum de rejeter H_0 à tort**.

➤ *Dans le cas d'un test unilatéral à gauche (H'_1)*



Si $t_c > -t_T$, c'est-à-dire si t_c se situe dans la zone de non rejet de H_0, alors on ne rejette pas l'hypothèse H_0. Et on conclut à une inégalité à gauche non significative entre les deux moyennes, au risque $\alpha = 5$ %, avec t tabulé, t_T, lu dans la table de t (pour $\alpha_t = 10$ %), avec un nombre de degré de liberté égal à (n - 1). On conclut alors que les deux séries appariées sont de moyennes équivalentes, c'est-à-dire de moyenne commune μ.

Si $t_c \leq -t_T$, c'est-à-dire si t_c se situe dans la zone de rejet de H_0, alors on rejette l'hypothèse H_0 et on conclut à une inégalité à gauche significative entre les deux moyennes, au risque $\alpha = 5$ %. Ainsi, la moyenne A est significativement inférieure à la moyenne B.

Le risque α', lu dans la table de t, correspondant à la valeur la plus proche de t_c mais inférieure (en valeur absolue), **fixe le degré de signification de l'inégalité** (on le note : $p < \alpha'$).

➤ *Dans le cas d'un test unilatéral à droite (H''_1)*

On peut se ramener facilement à un test unilatéral à gauche en inversant A et B, dans le calcul de m_d et s_d.

Sinon, si $t_c < +t_T$, on ne rejette pas l'hypothèse H_0, au risque $\alpha = 5$ %.

Puis, si $t_c \geq +t_T$, on rejette l'hypothèse H_0, au risque $\alpha = 5$ %. Et on conclut que la moyenne A est significativement supérieure à la moyenne B. Puis, de la même façon, le risque α' fixe le degré de signification de l'inégalité.

2.3. Comparaison de deux moyennes observées - test de Wilcoxon

2.3.1. Rôle du test

Ce test de Wilcoxon permet de comparer les lois observées de deux échantillons *appariés* (cf. glossaire), c'est-à-dire quand chaque sujet observé (ou unité d'observation) est son propre témoin :

- quand il existe une structure d'ordre entre les observations ;
- quand la différence entre les deux échantillons indique non seulement la direction mais également l'amplitude, sans toutefois permettre une "numérisation" complète de cette amplitude ;
- sans imposer de contrainte sur les distributions des deux séries appariées ou des différences.

Le test de Wilcoxon est aussi appelé **"test de signe et rang"**.

2.3.2. Procédure du test

Soit deux échantillons $x_1, \ldots, x_n$ de caractère X et de taille n
$x'_1, \ldots, x'_n$ de caractère X' et de taille n

Les deux échantillons sont appariés et de même taille.
Les hypothèses formulées peuvent être alors :

H_0 : X et X' sont de même loi
(i.e. : $F_X = F_{X'}$ $\forall x \in \mathbb{R}$ $F_X(x) = F_{X'}(x)$)

contre
- H_1 : X et X' ne sont pas de même loi (i.e. : $F_X \neq F_{X'}$)
- ou H_1' : $X > X'$ soit $F_X > F_{X'}$
- ou H_1'' : $X < X'$ soit $F_X < F_{X'}$

On note Δ_k la différence entre chaque observation couplée, soit $\Delta_k = x_k - x'_k$.

On considère la valeur absolue des Δ_k, soit $|\Delta_k|$, suivant leurs valeurs croissantes et on leur assigne un "rang" muni du signe + si $\Delta_k > 0$ ou du signe - si $\Delta_k < 0$. On élimine les Δ_k nuls ($\Delta_k = 0$).

• Exemple :

Pour $\Delta_1 = -2,1$ $\Delta_2 = 0,1$ $\Delta_3 = -4,7$

on classe les $|\Delta_k|$ dans l'ordre croissant :

|0,1|	|-2,1|	|-4,7|

et on leur assigne un rang conservant leur signe

+1	-2	-3

■ ***Problème des nœuds :***

Il existe deux sortes de nœuds
- si $\Delta_k = 0$ on élimine le couple (x_k, x'_k) correspondant ;
- si plusieurs $|\Delta_k|$ sont égaux, on attribue comme rang à chacun de ces termes $|\Delta_k|$ la moyenne arithmétique de leur rang naturel respectif. Comme précédemment on munit ce rang du signe + ou du signe -, selon que $\Delta_k > 0$ ou $\Delta_k < 0$.

• Exemple :

Pour	$\Delta_1 = -1,3$	$\Delta_2 = -1,7$	$\Delta_3 = -1,7$	$\Delta_4 = 2,3$
$\lvert\Delta_k\rvert$	|-1,3|	|-1,7|	|1,7|	|2,3|
rang	-1	-2,5	+2,5	4

Puis on note :

W_+ la somme des rangs de signe +

W_- la somme des rangs de signe -

W le plus petit des nombres W_+ et W_-, soit $W = \min(W_+, W_-)$

et $\tilde{W}$ le plus grand des nombres W_+ et W_-, soit $\tilde{W} = \max(W_+, W_-)$

Or, les deux variables aléatoires W et $\tilde{W}$ sont liées par la relation suivante :

soit $W + \tilde{W}$ = Somme des n premiers entiers

avec n : le nombre de paires non nulles $W + \tilde{W} = \frac{n(n+1)}{2}$

Les deux variables aléatoires W et $\tilde{W}$ ont la même loi de probabilité et donc la même espérance et le même écart-type :

$$E(W) = E(\tilde{W}) = \frac{n(n+1)}{4}$$

$$\sigma(W) = \sigma(\tilde{W}) = \sqrt{\frac{n(n+1)(2n+1)}{24}}$$

$$= \sqrt{var(w)}$$

On a la relation suivante qui permet de vérifier les calculs :

$$W + \tilde{W} = \frac{n(n+1)}{2}$$

Remarque :

Ce test n'est applicable que pour des échantillons de taille $n \geq 6$.

2.3.3. Test de Wilcoxon si échantillon de petite taille n ≤ 20, sans nœud

On utilise la table de Wilcoxon afin de déterminer **si on rejette ou non l'hypothèse nulle**, selon la valeur obtenue pour **le plus petit des deux totaux W_-** (somme des rangs des différences Δ_k de signe négatif) **ou W_+** (somme des rangs des différences Δ_k de signe positif).

On note W_C la plus petite des deux valeurs W_- ou W_+ (C correspond à calculé).

La table de Wilcoxon permet de déterminer la limite supérieure accordée à W_C pour conclure au rejet de H_0. S'il s'agit alors de W_+, on compare $W_C = W_+$ à la valeur donnée par la table de Wilcoxon, W_T (T pour tabulé), pour $\alpha = 5\ \%$. La limite supérieure accordée à W_C pour le rejet de H_0 est lue dans la table selon la taille des deux séries de données appariées. Par exemple, pour n = 20 on lit pour $\alpha = 5\ \%$, $W_T = 52$, et pour $\alpha = 1\ \%$, $W_T = 38$.

Autrement dit :

Si au seuil $\alpha = 5\ \%$, $W_C > W_T$ alors on conclut au non rejet de H_0.

Si au seuil $\alpha = 5\ \%$, $W_C \leq W_T$ alors on conclut au rejet de H_0. De même, si pour $\alpha = 1\ \%$, $W_C \leq W_T$ alors on conclut encore au rejet de H_0, et $\alpha = 1\ \%$ fixe le degré de signification de la différence entre les deux séries observées, soit $p < 0{,}01$.

▩ *Application :*

Voir l'exercice corrigé numéro 4 (chap. IX : p. 260 et 273).

2.3.4. Test de Wilcoxon si échantillon de petite taille $n \leq 20$, avec nœud

On procède comme dans le premier cas (voir 2.3.3.), mais en déterminant les rangs en tenant compte des nœuds.

Dans le cas de nœuds, on attribue comme rang à chacun de ces termes $|\Delta_k|$ égaux la moyenne arithmétique de leur rang naturel respectif en respectant leur signe.

2.3.5. Test de Wilcoxon si échantillon de grande taille $n > 20$, sans nœud

On utilise la table de la loi normale centrée réduite N(0, 1).

➤ *Dans le cas d'un test unilatéral à gauche*

On détermine la variable aléatoire normale ε calculée :

$$\varepsilon_c = \frac{W - \frac{n(n+1)}{4}}{\sqrt{\frac{n(n+1)(2n+1)}{24}}} = \frac{W - E(W)}{\sqrt{Var(W)}}$$

➤ *Dans le cas d'un test unilatéral à droite*

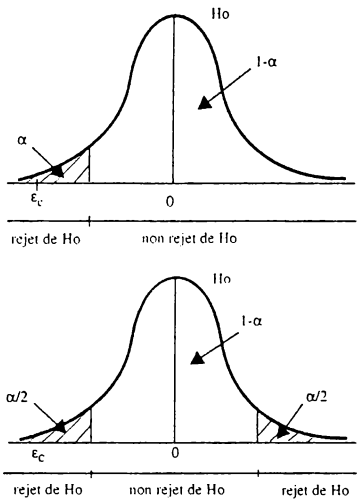
On détermine la variable aléatoire normale $\tilde{\varepsilon}$ calculée :

$$\tilde{\varepsilon}_c = \frac{\tilde{W} - \frac{n(n+1)}{4}}{\sqrt{\frac{n(n+1)(2n+1)}{24}}} = \frac{\tilde{W} - E(\tilde{W})}{\sqrt{Var(\tilde{W})}}$$

➤ *Dans le cas d'un test bilatéral*

On détermine l'une des deux variables aléatoires normales ε ou $\tilde{\varepsilon}$ calculée.

Puis au seuil α choisi, si ε ou $\tilde{\varepsilon}$ se situe dans la zone de rejet de H_0, on rejette l'hypothèse H_0.



▩ *Application :*
Voir l'exercice corrigé numéro 5 du chapitre X.

2.3.6. Test de Wilcoxon si échantillon de grande taille n > 20, avec nœud

On procède comme dans le troisième cas (voir 2.3.5.) en utilisant la table de la loi normale centrée réduite N(0, 1) et en déterminant les rangs des Δ_k, en tenant compte des nœuds.

3. Récapitulatif

Les tableaux suivants représentent un récapitulatif des tests à utiliser dans le cas de variables quantitatives, dans le cas de deux groupes.

Récapitulatif des tests à utiliser pour des variables quantitatives dans le cas de 2 groupes non appariés

Tests	Rôle du test	Type de variable	Population	Conditions d'applications
Test ε	Comparaison d'une moyenne observée à une moyenne théorique	Quantitative	Un échantillon observé comparé à une population théorique	$n \geq 30$
Test ε	Comparaison de deux moyennes observées m_A et m_B de taille respective n_A et n_B	Quantitative	Deux échantillons observés A et B indépendants	$n_A \geq 30$ et $n_B \geq 30$
Test t	Comparaison d'une moyenne observée à une moyenne théorique	Quantitative	Un échantillon observé comparé à une population héorique	variable quantitative observée distribuée normalement
Test t	Comparaison de deux moyennes observées m_A et m_B de taille respective n_A et n_B	Quantitative	Deux échantillons observés A et B indépendants	La variable quantitative étudiée est distribuée, dans les deux populations d'où proviennent les échantillons, selon des lois normales de même variance s^2.

Formule	Table	Conclusions
$\varepsilon_c = \dfrac{m-\mu}{\frac{s}{\sqrt{n}}}$ avec $s^2 = \dfrac{\sum (x_i - m)^2}{n-1}$	loi normale	si $\lvert\varepsilon_c\rvert \geq \varepsilon_T$ on rejette H_0 au seuil α
$\varepsilon_c = \dfrac{m_A - m_B}{\sqrt{\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}}}$ avec $s_A^2 = \dfrac{\sum (x_i - m_A)^2}{n_A - 1}$	loi normale	si $\lvert\varepsilon_c\rvert \geq \varepsilon_T$ on rejette H_0 au seuil α
$t_c = \dfrac{m-\mu}{\frac{s}{\sqrt{n}}}$ avec $s^2 = \dfrac{\sum (x_i - m)^2}{n-1}$	t	si $\lvert t_c\rvert \geq t_T$ on rejette H_0 au seuil α avec ddl = n - 1
$t_c = \dfrac{m_A - m_B}{\sqrt{\frac{s^2}{n_A} + \frac{s^2}{n_B}}}$ avec $s^2 = \dfrac{s^2 (n_A - 1) + s^2 (n_B - 1)}{n_A + n_B - 2}$	t	si $\lvert t_c\rvert \geq t_T$ on rejette H_0 au seuil α avec ddl = $n_A + n_B - 2$

Récapitulatif des tests à utiliser pour des variables quantitatives dans le cas de 2 groupes non appariés (suite)

Tests	Rôle du test	Type de variable	Population	Conditions d'applications
Mann et Whitney	Comparaison de deux moyennes observées ou comparaison de deux lois	Quantitative, semi-quantitative, ou qualitative avec une relation d'ordre	Deux échantillons indépendants	si n_1 et $n_2 \geq 9$ échantillons sans nœud
				si n_1 et $n_2 \geq 9$ échantillons avec nœud
				si n_1 et $n_2 \leq 8$ échantillons sans nœud
				si $n_1 \leq 8$ et $9 \leq n_2 \leq 20$ échantillons sans nœud
				si $n_1 \leq 8$ et $n_2 > 20$ échantillons sans nœud ou avec nœud
				si $n_1 \leq 8$ et $n_2 \leq 8$ échantillons avec nœud
F	Comparaison de deux variances	Quantitative	deux échantillons indépendants	les deux échantillons sont issus de populations pour lesquelles le caractère étudié est distribué normalement

Formule	Table	Conclusions
$V_i = \frac{U_i - E(U_i)}{\sqrt{V(U_i)}} = \frac{U_i - \frac{n_1 n_2}{2}}{\sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}}}$ avec U_1 = le nombre de couples (xi,x'j) tels que xi>x'j + 1/2 du nombre de couples ex-aequo et avec U_2 = le nombre de couples (xi,x'j) tels que xi<x'j + 1/2 du nombre de couples ex-aequo	loi normale	si $Vi \geq \varepsilon_T$ on rejette H_0 au seuil α
$V^* = \frac{U_i^* - \frac{n_1 n_2}{2}}{\sqrt{Var(U_i^*)}} = \frac{U_i^* - E(U_i^*)}{\sqrt{Var(U_i^*)}}$ avec $E(U_i^*) = \frac{n_1 n_2}{2}$ et $Var(U_i^*) = \frac{n_1 n_2 (N+1)}{12} - \frac{n_1 n_2 \sum_{i=1}^{e} (d_i^3 - d_i)}{12 N(N-1)}$	loi normale	si $V^* \geq \varepsilon_T$ on rejette H_0 au seuil
on prend la plus petite valeur entre U_1 et U_2	Mann et Whitney	si $U_c \leq U_T$ on rejette H_0 au seuil α
on prend la plus petite valeur entre U_1 et U_2	Mann et Whitney	si $U_c \leq U_T$ on rejette H_0 au seuil α
idem que dans le premier et le deuxième cas selon que l'on dispose ou non de nœud	loi normale	si $U_c \leq U_T$ on rejette H_0 au seuil α
on prend la plus petite valeur entre U_1 ou U_2	Mann et Whitney	si $U_c \leq U_T$ on rejette H_0 au seuil α
$F_c = \frac{S_A^2}{S_B^2}$	F	Si $F_c \geq F_T$ on rejette H_0 avec F_T lue à l'intersection de la colonne $(n_A\text{-}1)$ et de la ligne $(n_B\text{-}1)$ "à point 2,5 %"

Récapitulatif des tests à utiliser pour des variables quantitatives dans le cas de 2 groupes appariés

Tests	Rôle du test	Type de variable	Population	Conditions d'applications
ε	Comparaison de deux moyennes observées	Quantitative	Deux séries appariées	la taille des séries appariées, n, doit être supérieure ou égale à 30 ($n \geq 30$)
t	Comparaison de deux moyennes observées	Quantitative	Deux séries appariées	la différence entre les mesures doit être distribuée normalement
Wilcoxon	Comparaison de deux moyennes observées	Quantitative	Deux séries appariées	$n \geq 6$ dans tous les cas Si $n \leq 20$, sans nœud --- Si $n \leq 20$, avec nœud --- Si $n > 20$, sans nœud --- Si $n > 20$, avec nœud

Formule	Table	Conclusions
$\varepsilon_c = \dfrac{m_d}{\dfrac{s_d}{\sqrt{n}}}$ avec m_d la moyenne des différences et $s_d = \sqrt{\dfrac{\sum_{i=1}^{n}(d_i - m_d)^2}{n-1}}$	loi normale	si $\lvert\varepsilon_c\rvert \geq \varepsilon_T$ on rejette H_0 au seuil α
$tc = \dfrac{m_d}{\dfrac{s_d}{\sqrt{n}}}$ avec m_d la moyenne des différences et $s_d = \sqrt{\dfrac{\sum_{i=1}^{n}(d_i - m_d)^2}{n-1}}$	t	si $\lvert t_c\rvert \geq t_T$ on rejette H_0 au seuil α avec ddl = n - 1
On utilise le plus petit des deux totaux W- (somme des rangs des différences D_k de signe < 0) ou W+ (somme des rangs des différences D_k de signe > 0)	Wilcoxon	si au seuil α = 5 %, $W_C \leq W_T$ alors on conclut au rejet de H_0
Idem que dans le premier cas, mais en tenant compte des nœuds dans le calcul des rangs (pour les $\lvert\Delta_k\rvert$ égaux on calcule la moyenne arithmétique de leur rang naturel respectif en respectant leur signe)	Wilcoxon	si au seuil α = $W_C \leq W_T$ alors on conclut au rejet de H_0
$\varepsilon_c = \dfrac{W - \dfrac{n(n+1)}{4}}{\sqrt{\dfrac{n(n+1)(2n+1)}{24}}}$	loi normale	si $\lvert\varepsilon_c\rvert \geq \varepsilon_T$ on rejette H_0 au seuil α
Idem que dans le dernier cas, mais en tenant compte des nœuds dans le calcul des rangs (pour les $\lvert D_k\rvert$ égaux on calcule la moyenne arithmétique de leur rang	loi normale	si $\lvert\varepsilon_c\rvert \geq \varepsilon_T$ on rejette H_0 au seuil α

CHAPITRE VIII

Tests des différences entre variables quantitatives dans le cas de plus de deux groupes

1. **Analyse de variance à un facteur**
   Comparaison de plusieurs moyennes observées *225*
   1.1. Rôle du test *225*
   1.2. Principe du test *226*
   1.3. Procédure du test *229*
   1.4. Conditions d'application *231*

2. **Analyse de variance à deux facteurs croisés** *234*
   2.1. Rôle du test *234*
   2.2. Principe du test à partir d'un exemple *235*
   2.3. Procédure du test *237*
   2.4. Conditions d'application *240*

3. **Comparaison de plusieurs moyennes - Test de Kruskall-Wallis** *240*
   3.1. Rôle du test *240*
   3.2. Procédure du test *240*
   3.3. Cas des échantillons sans nœud *241*
   3.4. Cas des échantillons avec nœud *241*
   3.5. Exemples *242*

4. **Analyse de variance à deux facteurs avec répétition** *247*
   4.1. Rôle du test *247*
   4.2. Procédure du test *247*
   4.3. Conditions d'application *250*

5. **Récapitulatif** *251*

CHAPITRE VIII

Tests des différences entre variables quantitatives dans le cas de plus de deux groupes

Ce chapitre traite des différences entre variables quantitatives dans le cas de plus de deux groupes. Il présente notamment l'analyse de variance, c'est-à-dire la comparaison de plusieurs moyennes (plus de deux), dans le cas d'un facteur, puis de deux facteurs.

1. Analyse de variance à un facteur Comparaison de plusieurs moyennes observées

1.1. Rôle du test

L'analyse de variance consiste à comparer plusieurs moyennes, il s'agit d'une généralisation de la comparaison de deux moyennes. Disposant de plus de deux groupes, on se propose de savoir si, dans l'ensemble, ces groupes sont équivalents (c'est-à-dire si les moyennes du caractère étudié de ces groupes sont égales), ou si au contraire, ces groupes diffèrent dans l'ensemble (c'est-à-dire si au moins une de ces moyennes est différente des autres). Le terme "ANOVA" est parfois employé pour l'analyse de variance, il correspond à la contraction de la traduction en anglais "ANalyse Of VAriance".

Par exemple, dans le cas de la comparaison de trois moyennes, si les moyennes théoriques, μ_1, μ_2, et μ_3, des trois populations de référence de trois échantillons observés, ne sont pas toutes les mêmes, alors les moyennes observées à partir des trois échantillons étudiés auront davantage

tendance à être différentes que si les moyennes théoriques étaient équivalentes. **Il s'agit donc d'étudier la dispersion de plus de deux moyennes.** Si cette dispersion entre les moyennes est trop importante pour être expliquée par une fluctuation d'échantillonnage, alors on rejettera l'hypothèse nulle selon laquelle les moyennes sont équivalentes.

L'objectif est de déterminer si les échantillons observés sont issus ou non d'une même population de référence de moyenne commune μ pour le caractère étudié ***(hypothèse nulle)***. Si cela n'est pas le cas, on conclut que les moyennes observées diffèrent significativement dans leur ensemble, c'est-à-dire que les échantillons observés sont issus de populations de référence de moyennes $\mu_1, \mu_2, ..., \mu_n$, dont l'une au moins est différente des autres ***(hypothèse alternative)***.

■ <u>Les hypothèses</u>

Les hypothèses à tester sont :

H_0 : les moyennes du caractère étudié des n échantillons sont toutes égales.

H_1 : les moyennes du caractère étudié des n échantillons diffèrent dans leur ensemble (on observe au moins une différence entre les moyennes).

Ce qui revient à conclure si :

H_0 non rejeté : le facteur étudié n'est pas significatif.

H_1 non rejeté : le facteur étudié est significatif.

ou encore :

H_0 : $\mu_1, \mu_2, ..., \mu_n = \mu$

H_1 : $\mu_1, \mu_2, ..., \mu_n \neq \mu$ au moins pour une des moyennes

1.2. Principe du test

Il s'agit de comparer les moyennes μ_i d'une variable quantitative X, selon n groupes de sujets étudiés. Les données observées sont présentées dans un tableau tel que le tableau 1 suivant :

	Groupe 1	Groupe 2 ...	Groupe i ...	Groupe n
Valeurs observées de la variable X	x_{11} x_{12} x_{13} ... x_{1j} ...	x_{21} x_{22} x_{23} ... ... x_{2j} ...	x_{i1} x_{i2} x_{i3} ... ... x_{ij} ...	x_{n1} x_{n2} x_{n3} ... x_{nj} ...
Taille des groupes	n_1	n_2	n_j	n_n
Total	$\sum_{j=1}^{n_1} x_{1j}$	$\sum_{j=1}^{n_2} x_{2j}$	$\sum_{j=1}^{n_i} x_{ij}$	$\sum_{j=1}^{n_n} x_{nj}$
Moyennes observées	$m_1 = \frac{\sum_{j=1}^{n_1} x_{1j}}{n_1}$	$m_2 = \frac{\sum_{j=1}^{n_2} x_{2j}}{n_2}$	$m_i = \frac{\sum_{j=1}^{n_i} x_{ij}}{n_i}$	$m_n = \frac{\sum_{j=1}^{n_n} x_{nj}}{n_n}$

avec i désignant les colonnes et j les lignes du tableau.

Tableau 1

L'analyse de la variance est basée sur le calcul du rapport F (F en mémoire au statisticien anglais Sir Ronald Fisher (1890-1962)), entre la variance inter-colonnes (variance de la variable X étudiée entre les colonnes du tableau précédent) et la variance intra-colonne des observations (variance de la variable X étudiée à l'intérieur de chaque colonne du tableau précédent). Le rapport entre ces deux variances est noté F_c pour F calculé :

$$F_c = \frac{\text{variance inter-colonnes}}{\text{variance intra-colonne}}$$

$$\text{soit } F_c = \frac{\text{variabilité d'une mesure à l'autre entre les diverses colonnes}}{\text{variabilité des mesures dans une même colonne}}$$

$$= \frac{\text{variabilité d'une mesure d'une colonne à l'autre}}{\text{variabilité due au hasard (ou variance résiduelle)}}$$

Ces deux variances estimées à partir des observations doivent donc être proches, c'est-à-dire que **le rapport F doit être proche de 1** ; et l'on peut parier que F sera compris dans l'intervalle F_i à F_s, avec un risque d'erreur donné (risque d'erreur α lié à la fluctuation d'échantillonnage).

- ***La variance inter-colonnes*** (numérateur de F) peut être estimée avec un nombre de degré de liberté c - 1 (nombre de colonnes moins un, c'est-à-dire nombre de groupes étudiés moins un).

- ***La variance intra-colonne*** (dénominateur de F) peut être estimée avec un nombre de degré de liberté N - c (nombre d'observations moins le nombre de colonnes ; c'est-à-dire nombre de valeurs observées moins le nombre de groupes étudiés).

Les limites F_i et F_s sont alors lues dans la table de F à 5 % :

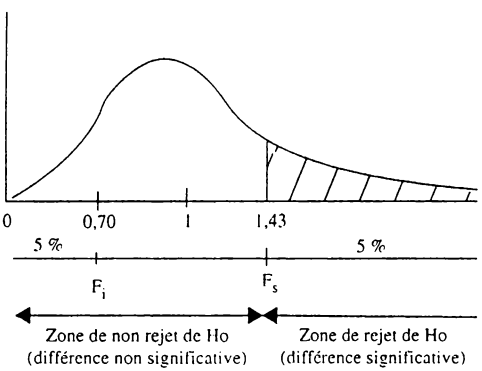
$$F_s = F_{N-c}^{c-1}$$

Dans la table de F "point 5 %", on lit F_s à l'intersection de la colonne c - 1 et de la ligne N - c.

Par exemple, $F_s = F_{120}^{60} = 1{,}43$

$$F_i = \frac{1}{F_{c-1}^{N-c}}$$

Par exemple, $F_i = \frac{1}{F_{60}^{120}} = \frac{1}{1{,}47} = 0{,}70$



Si $F_c \geq F_s$, la valeur de F_c est trop élevée pour être expliquée raisonnablement par les fluctuations d'échantillonnage, donc on rejette H_0 au risque 5% (différence significative).

Si F_c est compris entre F_i et F_s, rien n'empêche d'accepter H_0.

Si $F_c < F_i$, de même rien n'empêche d'accepter H_0, les différences observées sont liées aux fluctuations d'échantillonnage, donc rejet de H_1.

Il suffit alors de comparer uniquement la valeur de F_c à celle de F_s que l'on note plutôt F_T (F_T pour F tabulé). Avec :

$$F_T = F_{N-c}^{c-1}$$

1.3. Procédure du test

La comparaison des moyennes de c séries de mesures d'une quantité X, inscrites dans les c colonnes d'un tableau, est basée sur le calcul du rapport F_c :

F_c = variance entre-colonnes / variance résiduelle

$$F_c = \frac{\sum_{i=1}^{c}\left(\frac{T_i^2}{n_i}\right) - \frac{T_G^2}{N}}{c-1} \Big/ \frac{\sum_{i,j} x_{ij}^2 - \sum_{i=1}^{c}\left(\frac{T_i^2}{n_i}\right)}{N-c}$$

avec **i** et **j** correspondant respectivement aux colonnes et aux lignes ;

$\mathbf{n_i}$ = nombre de mesures de la colonne i

$\mathbf{N}$ = nombre total des mesures $= \sum_{i=1}^{c} n_i$

$\mathbf{T_i}$ = total des mesures de la colonne i $= \sum_{j=1}^{n_i} x_{ij}$

$\mathbf{T_G}$ = total général des mesures $= \sum_{i=1}^{c} T_i = \sum_{ij} x_{ij}$

$\mathbf{c}$ = nombre de séries de mesures = nombre de colonnes

$\sum_{i,j} x_{ij}^2$ = somme des carrés des mesures (de toutes les mesures)

$= \sum_{i=1}^{c}\sum_{j=1}^{n_i} x_{ij}^2$

Le calcul de F_c est effectué à partir des données observées réunies dans un tableau de la forme suivante :

Facteur A	**Groupe 1**	...	**Groupe i**	
Observations	x_{11} x_{12}	...	x_{i1} x_{i2} ... x_{ij}	
Somme des x (T_i)	$T_1 = \Sigma x_{1j}$	...	$T_i = \Sigma x_{ij}$	$\Sigma x = \Sigma T_i$
T_i^2	T_1^2	...		
T_i^2/n_i	T_1^2/n_1	...		$\Sigma(T_i^2/n_i)$
Somme des x^2	Σx_{1j}^2		Σx_{ij}^2	Σx^2

Tableau 2

On peut décomposer le calcul de F dans un tableau, dit tableau de l'analyse de variance (voir tableau 3 suivant), qui permet de vérifier les calculs.

Origine (1)	**Somme des carrés des écarts** (2)	**d.d.l.** (3)	**Variance** = (2)/(3)	**F_c** (rapport des deux variances)
Facteur A (variance entre colonnes) différences entre les moyennes	$\Sigma (T_i^2/n_i) - T_G^2/N$	c - 1	$\frac{\Sigma(T_i^2/n_i) - T_G^2/N}{c-1}$	$\frac{\Sigma(T_i^2/n_i) - T_G^2/N}{c-1}$ / $\frac{\Sigma x^2 - \Sigma(T_i^2/n_i)}{N-c}$
Résiduelle (variance intra-colonne) différences entre les observations et les moyennes	$\Sigma x^2 - \Sigma (T_i^2/n_i)$	N - c	$\frac{\Sigma x^2 - \Sigma(T_i^2/n_i)}{N-c}$	
Total	$\Sigma x^2 - T_G^2/N$	N - 1		

Conclusion : au seuil $\alpha = 5$ %, si F_c est supérieur ou égal à F_{N-c}^{c-1} lu dans la table de F à "point 5 %", alors on conclut que les moyennes du facteur A diffèrent significativement dans l'ensemble. Puis α lu dans les tables suivantes, "point 2,5 %" ..., avec les mêmes nombres de ddl, toujours pour un rejet de H_0, fixe le degré de signification (par exemple $p < 0{,}01$).

Tableau 3 : Tableau d'analyse de variance

■ *En résumé :*

• Si $F_c > F_T$ (avec $F_T = F_{N-c}^{c-1}$), les moyennes diffèrent significativement dans l'ensemble au risque 5 % (avec F_{N-c}^{c-1} lu dans la table de F "point 5 %" à l'intersection de la colonne c - 1 et de la ligne N - c) pour les nombres de degrés de liberté (c - 1) et (N - c).

• Si $F_c \leq F_T$ (avec $F_T = F_{N-c}^{c-1}$), les moyennes ne diffèrent pas significativement, elles sont équivalentes.

1.4. Conditions d'application

Le caractère étudié est distribué normalement dans chacune des populations de référence. De plus, le facteur étudié (groupe) est susceptible de modifier la valeur moyenne du caractère, mais pas sa variance. Autrement dit, le caractère étudié est de **même variance dans chacune des populations de référence** (ce qui demande de réaliser un test de comparaison des variances deux à deux ou global au préalable de l'analyse de variance).

■ *Remarque 1 :*

La taille des échantillons observés n'est pas obligatoirement la même.

■ *Remarque 2 :*

Le test est "robuste", c'est-à-dire encore valable lorsque les conditions d'application ne sont qu'à peu près vérifiées ; ceci d'autant plus que les effectifs des diverses séries sont plus grands et d'un même ordre de grandeur.

■ *Remarque 3 :*

Pour simplifier les calculs, on peut faire un changement d'origine, c'est-à-dire retrancher une constante K à toutes les valeurs observées (- K ou + K) ; et/ou faire un changement d'unité, c'est-à-dire multiplier toutes les valeurs observées par une constante K (K ou -K).

• Exemple :

Une échelle de mesure de satisfaction a permis de calculer le score de satisfaction de trois groupes de sujets comparables (toutes choses égales par ailleurs) concernant trois somnifères. Chaque échantillon est composé de douze sujets.

Les résultats sont présentés dans le tableau 4 ci-dessous :

	Somnifère 1	Somnifère 2	Somnifère 3	
	12	12	17	
	13	11	15	
	10	16	14	
	9	14	14	
Scores de	14	13	15	
satisfaction	12	13	12	
des sujets	11	12	11	
	8	16	16	
	10	14	13	
	10	13	12	
	12	13	17	
	14	15	15	
Somme des x (T_i)	135	162	171	$\Sigma x = 468$
T_i^2	18225	26244	29241	
T_i^2/n_i	1518,75	2187	2436,75	$\Sigma(T_i^2/n_i) = 6142,5$
Somme des x^2	1559	2214	2479	$\Sigma x^2 = 6252$

Tableau 4

1. En terme de satisfaction, y a-t-il une différence globale entre ces trois somnifères ?
2. En terme de satisfaction, y a-t-il une différence entre le somnifère 1 et le somnifère 2 ?

• *Correction de l'exemple :*

1. Les hypothèses correspondant à la question posée sont :

H_0 : Les moyennes des scores de satisfaction aux trois somnifères ne diffèrent pas significativement.
H_1 : Les moyennes des scores de satisfaction aux trois somnifères diffèrent significativement dans l'ensemble.

En supposant que le score de satisfaction soit distribué normalement et de même variance pour les trois échantillons, alors la comparaison de ces trois moyennes sera effectuée par une analyse de variance à un facteur.

Le tableau d'analyse de variance permettant de comparer globalement les trois somnifères est le suivant :

Origine (1)	Somme des carrés des écarts (2)	d.d.l. (3)	Variance = (2)/(3)	F_c
Entre produits (entre colonnes)	$\Sigma (T_i^2 / n_i) - T_G^2 / N$ $= 6142,5 - 468^2/36 = 58,5$	c-1 = 3-1 = 2	$\frac{58,5}{2} = 29,25$	$\frac{29,25}{3,32} = 8,81$
Résiduelle (intra-colonne)	$\Sigma x^2 - \Sigma (T_i^2 / n_i)$ $= 6252 - 6142,5 = 109,5$	N - c = 36-3=33	$\frac{109,5}{33} = 3,32$	
Total	$\Sigma x^2 - T_G^2 / N$ $= 6252 - 468^2/36 = 168$	N - 1 = 36 - 1 = 35		

Au seuil $\alpha = 5\ \%$ $F^2_{33} \approx F^2_{30} = 3,32$ donc $F_c = 8,81 > F^2_{33}$. Conclusion : en terme de satisfaction des sujets, les trois somnifères diffèrent significativement ($p < 0,001$).

Tableau 5 : Tableau d'analyse de variance

2. Pour comparer les moyennes théoriques des scores de satisfaction des sujets au somnifère 1 et au somnifère 2 (à partir des moyennes observées $m_1 = 11,25$ et $m_2 = 13,5$), on réalise un test de comparaison de deux moyennes (test t, la variance commune aux trois échantillons correspondant à la variance résiduelle). La question que l'on se pose est de savoir si les deux échantillons sont issus d'une même population de référence de moyenne commune m, ou si au contraire ils sont issus de deux populations différentes de moyennes théoriques différentes.

Les hypothèses à tester sont alors les suivantes :

$H_0 : \mu_1 = \mu_2 = \mu$
$H_1 : \mu_1 \neq \mu_2$

C'est-à-dire :

H_0 : La moyenne des scores de satisfaction du somnifère 1 est égale à celle du somnifère 2.
H_1 : La moyenne des scores de satisfaction du somnifère 1 diffère significativement de celle du somnifère 2.

On calcule alors :

$$t_c = \frac{|m_1 - m_2|}{\sqrt{\frac{s^2}{n_1} + \frac{s^2}{n_2}}}$$ avec s^2 correspondant à la variance résiduelle.

$$t_c = \frac{|11,25 - 13,5|}{\sqrt{\frac{3,32}{12} + \frac{3,32}{12}}} = \frac{2,25}{0,74} = 3,025$$

Le nombre de degrés de liberté du t "tabulé" est celui de la variance résiduelle, c'est-à-dire 33, soit $t_T = 2,042$; alors $t_c > t_T$.
Au risque $\alpha = 5$ %, on met alors en évidence une différence significative entre ces deux moyennes, c'est-à-dire une différence significative entre les somnifères 1 et 2, en terme de satisfaction des sujets. Les sujets sont satisfaits de façon significativement différente selon le somnifère 1 ou 2.

Cette différence de satisfaction n'est peut-être pas seulement attribuable aux somnifères, mais aux groupes eux-mêmes. En effet, on peut envisager qu'un des deux groupes soit plus facile à satisfaire au moment de l'étude, en raison par exemple de davantage de nuits blanches passées juste avant l'étude. Il est précisé que ces groupes sont comparables, ainsi cela suppose que les facteurs de confusion présumés soient contrôlés.

■ *Remarque :*

Nous insistons à nouveau sur l'attention et la prudence, toute particulière, qu'il faut mettre en œuvre lors de l'interprétation de résultats, afin de prendre en compte tout éventuel facteur de confusion.

2. Analyse de variance à deux facteurs croisés

2.1. Rôle du test

L'intérêt de l'analyse de variance à deux facteurs croisés est le même que celui de l'analyse de variance à un facteur, soit comparer des moyennes dans leur ensemble afin de **déterminer si les échantillons observés sont issus d'une même population de référence** ou non. Dans l'analyse de variance précédente, on étudiait la différence entre plusieurs groupes de sujets selon un facteur (facteur groupe : score de satisfaction différent selon le groupe). Cependant, on peut envisager des plans à plusieurs facteurs (exemples : facteur groupe et facteur période, ou facteur groupe et facteur traitement).

2.2. Principe du test à partir d'un exemple

On étudie le score de qualité de vie de 15 personnes atteintes d'un cancer du foie, avant, pendant, et après une période de traitement par chimiothérapie. On cherche à déterminer si les scores de qualité de vie sont les mêmes selon ces trois périodes de temps, ou si au contraire ils diffèrent significativement dans l'ensemble. On cherche également à déterminer si les scores de qualité de vie sont les mêmes selon les personnes, ou si au contraire ils diffèrent significativement dans l'ensemble. Notons que, dans cet exemple, plus le score est grand, plus la qualité de vie est bonne.

Il est nécessaire ici de tenir compte de la correspondance entre les données, soit entre les périodes, soit entre les sujets.

Les données sont les suivantes :

Personne	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Période															
Avant traitement	75	76	70	65	55	54	72	65	80	56	78	60	65	50	45
Pendant traitement	70	72	65	64	50	50	70	60	74	50	75	55	60	45	40
Après traitement	69	70	68	63	50	45	70	40	72	50	74	65	65	40	60

Afin de simplifier les calculs, il est souhaitable de diminuer les valeurs observées. C'est pourquoi, dans le tableau 6 suivant, les données sont diminuées de 61 (61 correspondant à une valeur proche de la moyenne générale de toutes les observations).

Bien entendu le changement d'origine et/ou d'unité n'est pas nécessaire lorsque l'on utilise les programmes des machines à calculer ou lorsque l'on utilise des logiciels statistiques ou des tableurs. Il ne modifie pas les résultats et n'est pas obligatoire.

Personne	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	T_i	T_i^2	n_i
Période																		
Avant Tmt	14	15	9	4	-6	-7	11	4	19	-5	17	-1	4	-11	-16	51	2601	15
Pendant Tmt	9	11	4	3	-11	-11	9	-1	13	-11	14	-6	-1	-16	-21	-15	225	15
Après Tmt	8	9	7	2	-11	-16	9	-21	11	-11	13	4	4	-21	-1	-14	196	15
T_j	31	35	20	9	-28	-34	29	-18	43	-27	44	-3	7	-48	-38	T_i=22	ΣT_i^2=3022	
T_j^2	961	1225	400	81	784	1156	841	324	1849	729	1936	9	49	2304	1444	ΣT_i^2=14092		
$\Sigma x^2{}_j$	341	427	146	29	278	426	283	458	651	267	654	53	33	818	698	Σx^2=5562		
n_j	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3			N=45

avec $\Sigma x^2{}_j$ correspondant à la somme des x^2 pour chaque colonne j.

Tableau 6

Données diminuées de 61 et calculs des paramètres de l'analyse de variance

Origine (1)	Somme des carrés des écarts (2)	d.d.l. (3)	Variance = (2)/(3)	F_c (F_{c1} et F_{c2})
Entre périodes	$\Sigma T_i^2/n_i - T_G^2/N$ $= 3022/15 - 22^2/45 = 190,71$	3 - 1 = 2	190,71/2 = 95,36	$\frac{95,35}{24,07} = 3,96$
Entre personnes	$\Sigma T_j^2/n_j - T_G^2/N$ $= 14092/3 - 22^2/45 = 4686,58$	15 - 1 = 14	4686,58/14 = 334,76	$\frac{334,76}{24,07} = 13,91$
Résiduelle : Total - (entre périodes + entre personnes)	5551,24 - (190,71+4686,58) = 673,95	44 - 16 = 28	673,95/28 = 24,07	
Total	$\Sigma x^2 - T_G^2/N$ $= 5562 - 22^2/45 = 5551,24$	45 – 1= 44		
A 5 ‰ $F_{28}^2 = 3,34$ donc $F_c = 4 > F_{28}^2$. Conclusion : le facteur période est significatif ($p < 0,05$). A 1 ‰ $F_{28}^{14} = 3,86$ donc $F_c = 14 > F_{28}^{14}$. Conclusion : le facteur personne est significatif ($p < 0,001$).				

Tableau 7

Tableau d'analyse de variance à deux facteurs, facteurs période et personne

La ligne résiduelle (quatrième ligne) est obtenue en retranchant de la ligne total, la somme des lignes entre périodes et entre personnes (lignes 2 et 3).

Ainsi il existe aussi bien un facteur personne qu'un facteur période. On peut donc conclure à une différence significative entre les moyennes des scores de qualité de vie, dans l'ensemble, avant, pendant et après le traitement. Donc, le score moyen de qualité de vie diffère au cours du temps en raison du traitement. De plus, on conclut à une différence significative entre les moyennes des scores de qualité de vie, dans l'ensemble, selon les personnes atteintes d'un cancer. Donc, le score moyen de qualité de vie diffère selon les patients, ce qui conduit à envisager une étude plus approfondie en tenant compte notamment du stade de cancer des sujets observés.

Les scores moyens de qualité de vie étant différents significativement dans l'ensemble selon les périodes de traitement, une comparaison des scores moyens par période deux à deux permettra de déterminer plus en détail où se situe cette différence. Le test t (compte tenu des effectifs) de comparaison de deux moyennes observées sera alors utilisé.

2.3. Procédure du test

Le tableau de données à analyser est de la forme suivante :

Facteur 2 / Facteur 1	modalité 1	modalité 2	...	modalité j	...	T_i	T_i^2	n_i
modalité 1	x_{11}	x_{12}	...	x_{1j}	...	$T_1 = \sum_{j=1}^{n_1} x_{1j}$	T_1^2	n_1
modalité 2	x_{21}	x_{22}	...	x_{2j}	...	$T_2 = \sum_{j=1}^{n_2} x_{2j}$	T_2^2	n_2
modalité i	x_{i1}	x_{i2}	...	x_{ij}	...		...	...
					...			
T_j	$T_1 = \sum_{i=1}^{n_1} x_{i1}$	$T_2 = \sum_{i=1}^{n_2} x_{i2}$			...	$T_G = \Sigma T_i = \Sigma T_j$ $= \Sigma x_{ij}$	ΣT_i^2	
T_j^2	T_1^2	$T_2{}^2$	...	...	...	ΣT_j^2		
Σx_j^2	$\sum_i x_{i1}^2$	$\sum_i x_{i2}^2$	...		...	$\Sigma x^2 = \sum_{ij} x_{ij}^2$		
n_j	n_1	n_2	...	...	...			$N = \Sigma n_i$ $= \Sigma n_j$

Tableau 8

Avec **i** et **j** correspondant respectivement aux lignes et aux colonnes

$\mathbf{x_{ij}}$: valeur observée à l'intersection de la ligne i et de la colonne j

$\mathbf{n_i}$: nombre de mesures de la ligne i

$\mathbf{n_j}$: nombre de mesures de la colonne j

Remarque :

Attention : ne pas confondre n_i lignes avec n_j colonnes ($n_{i.}$ et $n_{.j}$ seraient une meilleure notation, mais plus lourde).

N : nombre total des mesures $= \sum_{i=1}^{l} n_i = \sum_{j=1}^{c} n_j$

avec l : nombre de lignes, nombre de modalités du facteur 1

et c : nombre de colonnes, nombre de modalités du facteur 2

$\sum x^2_{.j}$: somme des x^2 sur chaque colonne j

T_i : total des mesures de la ligne i $= \sum_{j=1}^{n_i} x_{ij}$ (avec i fixé)

T_j : total des mesures de la colonne j $= \sum_{i=1}^{n_j} x_{ij}$ (avec j fixé)

T_G : total général des mesures $= \sum_{i=1}^{l} T_i = \sum_{j=1}^{c} T_j = \sum_{ij} x_{ij}$

$\sum_{i,j} x^2_{ij}$ = somme des carrés des mesures (de toutes les mesures)

$$= \sum_{i=1}^{l} \sum_{j=1}^{c} x^2_{ij} = \Sigma x^2 = \sum_{j} x^2_{.j} = \sum_{i} x^2_{i.}$$

Le tableau d'analyse de variance, dans le cas général, est alors le tableau 9 suivant. Nous avons noté par A, B, C, D, E, G, H, I, J, K, L et F, chaque cellule du tableau afin de mieux visualiser les calculs effectués.

Origine (1)	Somme des carrés des écarts (2)	d.d.l. (3)	Variance (2)/(3)	F_c
Entre facteur 1	$A = \Sigma T_i^2 / n_i - T_G^2 / N$	$E = \ell - 1$	$J = A/E$	$F_{c_1} = J/L$
Entre facteur 2	$B = \Sigma T_j^2 / n_j - T_G^2 / N$	$G = c - 1$	$K = B/G$	$F_{c_2} = K/L$
Résiduelle : Total - (Entre facteur 1 + Entre facteur 2)	$C = D - (A + B)$	$H = I - (E + G)$	$L = C/H$	
Total	$D = \Sigma x^2 - T_G^2 / N$	$I = N - 1$		

Tableau d'analyse de variance à deux facteurs
Tableau 9

■ **Lecture de F_H^E dans la table de F à "point 5 %"**

- Si $F_{c_1} > F_H^E$, on conclut que le facteur 1 est significatif, c'est-à-dire que les moyennes du caractère étudié diffèrent significativement, dans l'ensemble, selon le facteur 1. On recherche alors le *degré de signification* de cette différence, *c'est-à-dire la valeur du risque* α' *jusqu'à laquelle on conclut encore à une différence significative* ($p < \alpha'$).

- Si $F_{c_1} \leq F_H^E$, on conclut que le facteur 1 n'est pas significatif, c'est-à-dire que les moyennes du caractère étudié ne diffèrent pas significativement selon le facteur 1.

■ **Lecture de F_H^G dans la table de F à "point 5 %"**

- Si $F_{c_2} > F_H^G$, on conclut que le facteur 2 est significatif, c'est-à-dire que les moyennes du caractère étudié diffèrent significativement, dans l'ensemble, selon le facteur 2. On recherche alors le degré de signification de cette différence ($p < \alpha'$).

- Si $F_{c_2} \leq F_H^G$, on conclut que le facteur 2 n'est pas significatif, c'est-à-dire que les moyennes du caractère étudié ne diffèrent pas significativement selon le facteur 2.

Ainsi on doit comparer :

$$F_{c_1} = \frac{\dfrac{\sum_i T_i^2 / n_i - T_G^2 / N}{\ell - 1}}{\dfrac{\sum x^2 - \left(\sum_i T_i^2 / n_i - 2T_G^2 / N + \sum_i T_j^2 / n_j \right)}{N - \ell + c - 1}} \quad \text{à} \quad F_{N-\ell+c-1}^{\ell-1}$$

et

$$F_{c_2} = \frac{\dfrac{\sum_j T_j^2 / n_j - T_G^2 / N}{c - 1}}{\dfrac{\sum x^2 - \left(\sum_i T_i^2 / n_i - 2T_G^2 / N + \sum_i T_j^2 / n_j \right)}{N - \ell + c - 1}} \quad \text{à} \quad F_{N-\ell+c-1}^{c-1}$$

2.4. Conditions d'application

Les données observées doivent être **conformes à un modèle additif,** c'est-à-dire pour toutes cellules ij du tableau de données correspondant au facteur 1 et au facteur 2 : chaque valeur observée doit être proche de la moyenne générale des observations ; additionnée d'une contribution particulière de la ligne i du facteur 1 (une diminution ou une augmentation), qui doit être du même ordre pour toutes les colonnes j du facteur 2 ; additionnée d'une contribution particulière de la colonne j du facteur 2 (une diminution ou une augmentation), qui doit être du même ordre pour toutes les lignes i du facteur 1 ; et additionnée de la variation aléatoire de la mesure qui doit être distribuée selon une loi normale centrée, de variance égale à la variance résiduelle.

Le test est relativement "robuste", c'est-à-dire encore valable lorsque ces conditions ne sont vérifiées qu'approximativement.

3. Comparaison de plusieurs moyennes - Test de Kruskall-Wallis

3.1. Rôle du test

Comparaison de n caractères $X^{(1)}, X^{(2)}, \ldots, X^{(n)}$ de n échantillons *indépendants.* Le test de Kruskall-Wallis permet de réaliser une analyse de variance (comparaison de plusieurs moyennes) à un facteur sans contrainte sur la forme des distributions observées.

3.2. Procédure du test

Soit n échantillons :

$$\begin{cases} x_1^{(1)}, x_2^{(1)}, \ldots, x_{N_1}^{(1)} \text{ de taille } N_1, \text{ de caractère } X^{(1)} \\ x_1^{(2)}, x_2^{(2)}, \ldots, x_{N_2}^{(2)} \text{ de taille } N_2, \text{ de caractère } X^{(2)} \\ \\ \\ x_1^{(n)}, x_2^{(n)}, \ldots, x_{N_n}^{(n)} \text{ de taille } N_n, \text{ de caractère } X^{(n)} \end{cases}$$

Les variables aléatoires $X^{(1)}, X^{(2)}, \ldots, X^{(n)}$ sont *indépendantes.*

Les hypothèses formulées sont :

H_0 : les variables aléatoires $X^{(1)}, X^{(2)}, ..., X^{(n)}$ sont de même loi.
H_1 : les variables aléatoires $X^{(1)}, X^{(2)}, ..., X^{(n)}$ sont de loi différente.

ou encore

H_0 : les variables aléatoires $X^{(1)}, X^{(2)}, ..., X^{(n)}$ sont de même moyenne.
H_1 : les variables aléatoires $X^{(1)}, X^{(2)}, ..., X^{(n)}$ sont de moyenne différente.

On rassemble les termes des n échantillons en un seul échantillon global, en ordonnant les termes de cet échantillon dans l'ordre croissant.

3.3. Cas des échantillons sans nœud

On retient pour chaque terme de l'échantillon global ordonné, son rang.
On note R_j = Somme des rangs des termes du $j^{ième}$ échantillon (avec $1 \leq j \leq n$).

On calcule $$H = \frac{12}{N(N+1)} \sum_{j=1}^{n} \frac{R_j^2}{N_j} - 3(N+1)$$

avec $N = N_1 + N_2 + ... + N_n$ qui correspond à la taille de l'échantillon global.

Deux cas sont proposés ici :

1er cas : on dispose de trois échantillons (n = 3) de petites tailles avec au moins un des $N_j \leq 5$, dans ce cas on utilise la table de Kruskall Wallis.

2e cas : on dispose de trois échantillons (n = 3) de grandes tailles avec tous les $N_j > 5$, ou de plus de 4 échantillons (n > 4) de grandes tailles avec tous les $N_j > 5$, dans ces cas on utilise la table du χ^2 à n-1 degrés de liberté, car dans ces cas la variable H suit une loi du χ^2 à n-1 ddl.

3.4. Cas des échantillons avec nœud

On attribue à chacun des termes égaux de l'échantillon global ordonné, le rang égal à la moyenne arithmétique des rangs naturels des termes.
On note R_j^* = Somme des rangs des termes du $j^{ième}$ échantillon (avec $1 \leq j \leq n$).

On calcule $$H^* = \frac{\frac{12}{N(N+1)} \sum_{j=1}^{n} \frac{R_j^{*2}}{N_j} - 3(N+1)}{1 - \frac{\sum_{i=1}^{e} (d_i^3 - d_i)}{N^3 - N}}$$

avec e : le nombre des séries de termes égaux

avec $N = N_1 + N_2 + ... + N_n$

et avec d_i = le nombre de termes ayant la $i^{ème}$ valeur.

■ **Trois cas sont proposés ici :**

3e cas : on dispose de trois échantillons (n = 3) de petites tailles avec tous les Nj ≤ 5, dans ce cas on utilise la table de Kruskall Wallis avec H^*.

4e cas : si le nombre de termes égaux des différents nœuds de l'échantillon global excède 25 % de l'effectif de l'échantillon global, on utilise la table du χ^2 avec H^*, car dans ce cas H^* suit une loi du χ^2 à n-1 ddl.

5e cas : si le nombre de termes égaux des différents nœuds de l'échantillon global est inférieur à 25 % de l'effectif de l'échantillon global, on utilise la table du χ^2 avec H, car dans ce cas H suit une loi du χ^2 à n-1 ddl.

Le calcul du pourcentage de termes égaux des différents nœuds de l'échantillon global est obtenu par :

$$\frac{\text{nombre de termes avec noeuds}}{N}$$

avec N correspondant à la taille de l'échantillon global.

3.5. Exemples

• Exemple 1 :

Un chercheur désire comparer l'autoritarisme des enseignants administrateurs et non administrateurs dans les écoles. Pour cette étude il dispose de trois types de sujets :

Groupe 1 : les enseignants désirant rester enseignants.

Groupe 2 : les enseignants voulant être administrateurs.

Groupe 3 : les enseignants déjà administrateurs.

Il soumet un questionnaire à 14 sujets qui permet de leur attribuer une note d'autoritarisme. Les données observées sont les suivantes :

Groupe 1		Groupe 2		Groupe 3	
note	**rang**	**note**	**rang**	**note**	**rang**
96	4	82	2	115	7
128	9	124	8	149	13
83	3	132	10	166	14
61	1	135	11	147	12
101	5	109	6		

Poser les hypothèses et tester ces hypothèses au seuil α = 5 %.

Les hypothèses à tester sont :

$$\text{contre} \begin{cases} H_0 : \text{les 3 caractères ont même loi.} \\ H_1 : \text{les 3 caractères ont des lois différentes.} \end{cases}$$

soit

$$\text{contre} \begin{cases} H_0 : \text{les 3 groupes sont équivalents en terme d'autoritarisme.} \\ H_1 : \text{les 3 groupes diffèrent en terme d'autoritarisme.} \end{cases}$$

Un test de Kruskall Wallis dans le cas d'échantillons sans nœud de petites tailles permettra de tester ces hypothèses.

Calcul de la somme des rangs de chaque échantillon :

$R_1 = 22 \qquad R_2 = 37 \qquad R_3 = 46$

Puis on calcule la variable H :

$$H = \frac{12}{N(N+1)} \sum_{j=1}^{n} \frac{R_j^2}{N_j} - 3(N+1)$$

soit $$H = \frac{12}{14 \times 15}\left(\frac{(22)^2}{5} + \frac{(37)^2}{5} + \frac{(46)^2}{4}\right) - 3 \times 15$$

$H = 6{,}4$

On lit la valeur du paramètre tabulé dans la table de Kruskall Wallis, au seuil $\alpha = 5\ \%$, pour $n_1 = 5$, $n_2 = 5$ et $n_3 = 4$, soit $H_T = 5{,}6429$.

Ainsi $H_C > H_T$ donc on rejette l'hypothèse H_0 et on conclut que la différence entre les trois groupes en terme d'autoritarisme est significative avec un degré de signification $p < 0{,}049$ (en effet, au seuil $\alpha = 4{,}9\ \%$, on a encore $H_C > H_T$, puisque $H_T = 5{,}6657$).

Parmi les sujets observés, on peut donc dire que l'autoritarisme diffère selon le type d'enseignant (les enseignants désirant rester enseignants, les enseignants voulant être administrateurs, et les enseignants déjà administrateurs).

• *Exemple 2 :*

Trois méthodes d'enseignement des statistiques sont appliquées auprès de trois groupes d'étudiants de première année universitaire. Les résultats obtenus à un examen, après un semestre d'enseignement, sont les suivants :

Méthode A Groupe 1		Méthode B Groupe 2		Méthode C Groupe 3	
note	rang	note	rang	note	rang
60	1	80	10	74	7
69	4	72	5	86	14
76	8	61	2	87	15
73	6	84	12	91	17
67	3	79	9	94	18
89	16	82	11	97	19
		85	13		

Ces trois méthodes sont-elles équivalentes ?
Poser les hypothèses et tester ces hypothèses au seuil $\alpha = 5\ \%$.

Les hypothèses à tester sont :

$$\text{contre} \begin{cases} H_0 : \text{les trois méthodes sont équivalentes} \\ H_1 : \text{les trois méthodes sont différentes dans l'ensemble} \end{cases}$$

Un test de Kruskall Wallis dans le cas d'échantillons sans nœud de grandes tailles permettra de tester ces hypothèses.

Calcul de la somme des rangs de chaque échantillon :

$R_1 = 38 \quad R_2 = 62 \quad R_3 = 90$

Puis on calcule la variable H :

$$H = \frac{12}{N(N+1)} \sum_{j=1}^{n} \frac{R_j^2}{N_j} - 3(N+1)$$

soit $$H = \frac{12}{19 \times 20}\left(\frac{(38)^2}{6} + \frac{(62)^2}{7} + \frac{(90)^2}{6}\right) - 3 \times 20$$

$$H = 7{,}57$$

On lit la valeur du paramètre tabulé dans la table du χ^2 pour $\alpha = 5\ \%$ et ddl = n - 1 = 2, soit $H_T = 5{,}991$.
Ainsi $H_C > H_T$ donc on rejette l'hypothèse H_0 et on conclut que la différence entre les trois séries de notes est significative ($p < 0{,}05$). Il semble donc que certaines méthodes soient plus efficaces que d'autres, une étude comparative des méthodes deux à deux permettrait de distinguer la ou les deux meilleurs méthodes d'enseignement. Une analyse des résultats aux examens de statistique (selon ces trois méthodes d'enseignement), effectuée sur une plus large population, avec tirage au sort pour la constitution des trois groupes, permettrait de confirmer si la différence des notes obtenues est attribuable aux méthodes d'enseignement (et non à d'éventuels facteurs de confusion).

• *Exemple 3 :*

On se demande s'il existe une différence entre les pertes de poids (pds) consécutives à huit régimes, après une semaine, chez huit groupes de femmes de même âge. Les données observées (en kilogrammes) et le rang (rg) de chaque observation sont reportés dans le tableau 10 suivant.

Groupe 1 $X^{(1)}$		Groupe 2 $X^{(2)}$		Groupe 3 $X^{(3)}$		Groupe 4 $X^{(4)}$		Groupe 5 $X^{(5)}$		Groupe 6 $X^{(6)}$		Groupe 7 $X^{(7)}$		Groupe 8 $X^{(8)}$	
Pds	Rg	Pds	Rg	Pds	Rg	Pds	Rg	Pds	Rg	Pds	Rg	Pds	Rg	Pds	Rg
1	8,5	1,75	52,5	1.65	47,5	1.6	41	1,3	23	1,55	36	1,3	23	1.25	18.5
1.4	27,5	1.4	27,5	1.8	54,5	1,65	47.5	1,3	23	1.45	31.5	1.1	12.5	1.2	15.5
1.65	47.5	1.6	41	1.3	23	1.6	41	1.45	31.5	1.55	36	1.1	12.5	1.5	34
1.6	41	1.75	52.5	1.55	36	1.45	31,5	1	8.5	1.25	18.5	1.25	18.5	0.75	4
2.2	56	1.15	14	1.6	41	1.65	47.5	1	8.5			0.6	2.5		
1.8	54.5	1.2	15,5	1.65	47.5	1.25	18.5	1.05	11			0.6	2.5		
0.95	6	1	8.5	1.45	31,5	1.3	23								
1.65	47,5	0.8	5	1.7	51	1.4	27.5								
1.4	27,5			1.6	41										
0.55	1			1.6	41										

Tableau 10

Poser les hypothèses et tester ces hypothèses au seuil $\alpha = 5\ \%$.

Les hypothèses à tester sont :

$$\begin{cases} H_0 : \text{les pertes de poids selon les huit régimes sont équivalentes} \\ \text{contre} \\ H_1 : \text{les pertes de poids selon les huit régimes sont différentes} \end{cases}$$

Un test de Kruskall Wallis dans le cas d'échantillons avec nœud permet de tester ces hypothèses.

Taille des échantillons :

$N_1 = 10 \quad N_2 = 8 \quad N_3 = 10 \quad N_4 = 8 \quad N_5 = 6 \quad N_6 = 4 \quad N_7 = 6 \quad N_8 = 4$

donc $N = 56$

Le nombre de termes égaux des différents nœuds dans l'échantillon global est supérieur à 25 % :

$$\frac{\text{nombre de termes avec nœuds}}{N} = \frac{47}{56} = 0{,}8393 = 83{,}93\ \%$$

En effet, 47 termes se retrouvent plus d'une fois.

On calcule alors la variable H^* :

$$H^* = \frac{\dfrac{12}{N(N+1)} \sum_{j=1}^{n} \dfrac{R_j^{*2}}{N_j} - 3(N+1)}{1 - \dfrac{\sum_{i=1}^{e} (d_i^3 - d_i)}{N^3 - N}}$$

Pour cela on calcule la somme des rangs de chaque échantillon :
$R_1 = 317$ $R_2 = 216{,}5$ $R_3 = 414{,}5$ $R_4 = 277{,}5$ $R_5 = 105{,}5$ $R_6 = 122$ $R_7 = 71{,}5$ $R_8 = 72$

Puis on calcule les $d_i^3 - d_i$, avec d_i, le nombre de termes ayant la $i^{ème}$ valeur :

$d_1^3 - d_1 = 8 - 2 = 6$
$d_2^3 - d_2 = 64 - 4 = 60$
$d_3^3 - d_3 = 8 - 2 = 6$
$d_4^3 - d_4 = 64 - 4 = 60$
$d_5^3 - d_5 = 75 - 5 = 70$
$d_6^3 - d_6 = 64 - 4 = 60$
$d_7^3 - d_7 = 64 - 4 = 60$
$d_8^3 - d_8 = 27 - 3 = 24$
$d_9^3 - d_9 = 8 - 2 = 6$
$d_{10}^3 - d_{10} = 343 - 7 = 336$
$d_{11}^3 - d_{11} = 216 - 6 = 210$
$d_{12}^3 - d_{12} = 8 - 2 = 6$
$d_{13}^3 - d_{13} = 8 - 2 = 6$

Alors $\sum_{i=1}^{13} (d_i^3 - d_i) = 910$

Ainsi la variable H^* équivaut à :

$$H^* = \frac{\frac{12}{N(N+1)} \sum_{j=1}^{n} \frac{R_j^{*2}}{N_j} - 3(N+1)}{1 - \frac{\sum_{i=1}^{e} (d_i^3 - d_i)}{N^3 - N}}$$

soit

$$H^* = \frac{\frac{12}{56(57)} \left(\frac{317^2}{10} + \frac{216{,}5^2}{8} + \frac{414{,}5^2}{10} + \frac{277{,}5^2}{8} + \frac{105{,}5^2}{6} + \frac{122^2}{4} + \frac{71{,}5^2}{6} + \frac{72^2}{4} \right) - 3(56+1)}{1 - \frac{910}{56^3 - 56}}$$

$H^* = 18{,}56$

On lit la valeur du paramètre tabulé dans la table du χ^2 pour $\alpha = 5\ \%$ et ddl = n - 1 = 7, soit $H_T = 14{,}067$.
Ainsi $H_C > H_T$ donc on rejette l'hypothèse H_0 et on conclut que la différence des pertes de poids est significative ($p < 0{,}01$).

Il semble donc que certains régimes soient plus efficaces que d'autres, une étude comparative des régimes deux à deux permettrait de distinguer les meilleurs. Une étude identique, mais effectuée sur une plus large population, avec tirage au sort pour la constitution des huit groupes, permettrait de confirmer si la différence des pertes de poids observée est attribuable aux régimes (et non à d'éventuels facteurs de confusion).

4. Analyse de variance à deux facteurs avec répétition

4.1. Rôle du test

L'intérêt de l'analyse de variance à deux facteurs avec répétition est le même que celui de l'analyse de variance à un facteur ou à deux facteurs sans répétition, soit **comparer des moyennes dans leur ensemble** afin de déterminer si les échantillons observés sont issus d'une même population de référence ou non. Dans le cas présent **on envisage des plans à deux facteurs A et B** (par exemple : facteur groupe et facteur traitement) pour une mesure donnée (variable dépendante quantitative) qui peut être répétée plusieurs fois, k fois (avec k = 1, 2, ..., r).

4.2. Procédure du test

Les hypothèses à tester sont :

$$\begin{cases} H_0 : \begin{cases} \text{pas d'effet du facteur A} \\ \text{pas d'effet du facteur B} \\ \text{pas d'interaction entre A et B} \end{cases} \\ \text{contre } H_1 : \begin{cases} \text{effet de A} \\ \text{effet de B} \\ \text{interaction entre A et B} \end{cases} \end{cases}$$

Le tableau de données à analyser est de la forme suivante :

Facteur B / Facteur A	modalité 1	modalité 2	...	modalité j	...	modalité K_B	L_i	L_i^2	n_i
modalité 1	x_{11k}	x_{12k}	...	x_{1jk}	...		$L_1 = \sum_{j,k} x_{1jk}$	L_1^2	n_1
modalité 2	x_{21k}	x_{22k}	...	x_{2jk}	...		$L_2 = \sum_{j,k} x_{2jk}$	L_2^2	n_2
							...	...	...
modalité i	x_{i1k}	x_{i2k}	...	x_{ijk}	...		L_i	L_i^2	n_i
		...		...			...	...	...
modalité K_A			...		...		L_{KA}	L_{KA}^2	n_{KA}
C_j	$C_1 = \sum_{i,k} x_{i1k}$	$C_2 = \sum_{i,k} x_{i2k}$					$T_G = \Sigma L_i$ $= \Sigma C_j = \Sigma x_{ijk}$	ΣL_i^2	
C_j^2	C_1^2	C_2^2					ΣC_j^2		
$\Sigma_j x^2$	$\sum_{i,k} x_{i1k}^2$	$\sum_{i,k} x_{i2k}^2$					$\Sigma x^2 = \sum_{ijk} x_{ijk}^2$		
n_j	n_1	n_2							$N = \Sigma n_i$ $= \Sigma n_j$

Tableau 11

Avec **i** et **j** correspondant respectivement aux lignes et aux colonnes du tableau

k correspondant à la troisième dimension du tableau, soit les répétitions

$\mathbf{x_{ijk}}$: valeur observée à l'intersection de la ligne i et de la colonne j pour la k^e répétition

$\mathbf{n_i}$: nombre de mesures de la ligne i (comprenant les répétitions)

$\mathbf{n_j}$: nombre de mesures de la colonne j (comprenant les répétitions)

[voir remarque page 237]

N : nombre total des mesures = $\sum_{i=1}^{K_A} n_i = \sum_{j=1}^{K_B} n_j$

avec $\mathbf{K_A}$: le nombre de lignes = nombre de modalités du facteur A

et $\mathbf{K_B}$: le nombre de colonnes = nombre de modalités du facteur B

$\mathbf{L_i}$: somme des mesures de la ligne i = $\sum_{j,k} x_{ijk}$ (avec i fixé)

C_j : somme des mesures de la colonne j = $\sum_{i,k} x_{ijk}$ (avec j fixé)

T_{ij} : somme par case (i,j) = $\sum_{k} x_{ijk}$ (avec i et j fixés)

T_G : somme générale des mesures = $\sum_{i=1}^{K_A} L_i = \sum_{j=1}^{K_B} C_j = \sum_{ijk} x_{ijk}$

$\sum_{i,j,k} x_{ijk}^2$: somme des carrés de toutes les mesures = $\sum_i \sum_j \sum_k x_{ijk}^2 = \Sigma x^2$

Le tableau d'analyse de variance, dans le cas général, est alors le tableau suivant :

Origine	Somme des carrés (sc)	d.d.l.	Variance	F_c
Facteur A	$SC_A = \Sigma L_i^2 / (K_B.r) - T_G^2 / N$	$K_A - 1$	$S_A^2 = SC_A/ddl$	$F_A = S_A^2 / S_R^2$
Facteur B	$SC_B = \Sigma C_j^2 / (K_A.r) - T_G^2 / N$	$K_B - 1$	$S_B^2 = SC_B/ddl$	$F_B = S_B^2 / S_R^2$
Interaction entre A et B	$SC_{AxB} = SC_{inter\ cases} - (SC_A + SC_B)$	$(K_A - 1)(K_B - 1)$	$S_{AB}^2 = SC_{AB}/ddl$	$F_{AB} = S_{AB}^2 / S_R^2$
Entre cases	$SC_{inter\ cases} = \Sigma T_{ij}^2 / r - T_G^2 / N$	$K_A.K_B - 1$		
Résiduelle (R)	$SC_R = SC_T - SC_{inter\ cases}$	$K_A.K_B (r - 1)$	$S_R^2 = SC_R/ddl$	
Totale	$SC_T = \Sigma x^2 - T_G^2 / N$	$N - 1$		

Tableau 12
Tableau d'analyse de variance à deux facteurs avec répétition

■ Lecture dans la table de F à "point 5 %"

- Si $F_A > F^{K_A - 1}_{K_A K_B (r-1)}$ à 5 %, on conclut que le facteur A est significatif, c'est-à-dire que les moyennes du caractère étudié diffèrent significativement, dans l'ensemble, selon le facteur A. On recherche alors le degré de signification de cette différence, c'est-à-dire la valeur du risque α' jusqu'à laquelle on conclut encore à une différence significative ($p < \alpha'$).

- Si $F_B > F^{K_B-1}_{K_A K_B(r-1)}$ à 5 %, on conclut que le facteur B est significatif, c'est-à-dire que les moyennes du caractère étudié diffèrent significativement, dans l'ensemble, selon le facteur B. On recherche alors le degré de signification de cette différence, c'est-à-dire la valeur du risque α jusqu'à laquelle on conclut encore à une différence significative ($p < \alpha'$).

- Si $F_{AB} > F^{(K_A-1)(K_B-1)}_{K_A K_B(r-1)}$ à 5 %, on conclut que l'interaction entre A et B est significative, c'est-à-dire que les moyennes du caractère étudié diffèrent significativement, dans l'ensemble, selon l'interaction entre le facteur A et le facteur B. On recherche alors le degré de signification de cette différence, c'est-à-dire la valeur du risque α jusqu'à laquelle on conclut encore à une différence significative ($p < \alpha'$).Si l'interaction est significative, les deux tests précédents et leur interprétation doivent être modifiées. Si l'interaction est significative, les deux tests précédents et leur interprétation doivent être modifiés.

4.3. Conditions d'application

La variable aléatoire X étudiée suit une loi normale quelque soient les classes i et j fixées pour A et B.

La variabilité intracase est constante : $\sigma^2_{ij} = \sigma^2 \quad (\forall i, j)$.

5. Récapitulatif

Tableau 13 : Tests à utiliser lors de la comparaison de plus de deux moyennes, dans le cas d'un ou deux facteurs
(voir pages 252-253 et 254-255 suivantes).

Récapitulatif des tests à utiliser pour des variables quantitatives dans le cas de plus de 2 groupes

Tests	Rôle du test	Type de variable	Population	Conditions d'applications
Analyse de variance à un facteur	Comparaison de plusieurs moyennes observées (plus de 2)	Quantitative	n échantillons indépendants étudiés selon un facteur	le caractère étudié doit être distribué normalement et doit être de même variance dans chacune des populations de référence
Analyse de variance à deux facteurs	Comparaison de plusieurs moyennes observées (plus de 2)	Quantitative	n échantillons indépendants étudiés selon deux facteurs	les données observées doivent être conformes à un modèle additif

Tableau 13

Formule	Table	Conclusions
$F_c = \dfrac{\text{variance entre colonnes}}{\text{variance résiduelle}}$ $F_c = \dfrac{\dfrac{\sum_{i=1}^{c}\left(\dfrac{T_i^2}{n_i}\right) - \dfrac{T_G^2}{N}}{c-1}}{\dfrac{\sum_{i,j} x_{ij}^2 - \sum_{i=1}^{c}\left(\dfrac{T_i^2}{n_i}\right)}{N-c}}$	F	si $F_c \geq F_{N-c}^{c-1}$ le facteur étudié est significatif au seuil α avec F_T lu dans la table de F "point 5 %" à l'intersection de la colonne c-1 et de la ligne N-c
$F_{c1} = \dfrac{\dfrac{\sum_i T_i^2/n_i - T_G^2/N}{\ell - 1}}{\dfrac{\sum x^2 - \left(\sum_i T_i^2/n_i - 2T_G^2/N + \sum_i T_j^2/n_j\right)}{N-\ell+c-1}}$ et $F_{c2} = \dfrac{\dfrac{\sum_j T_j^2/n_j - T_G^2/N}{c-1}}{\dfrac{\sum x^2 - \left(\sum_i T_i^2/n_i - 2T_G^2/N + \sum_i T_j^2/n_j\right)}{N-\ell+c-1}}$	F	si $F_{c1} \geq F_{N-\ell+c-1}^{\ell-1}$ le facteur 1 est significatif au seuil α avec F_{T1} lu dans la table de F "point 5 %" à l'intersection de la colonne l-1 et de la ligne N-l+c-1 si $F_{c2} \geq F_{N-\ell+c-1}^{c-1}$ le facteur 2 est significatif au seuil α avec F_{T2} lu dans la table de F "point 5 %" à l'intersection de la colonne c-1 et de la ligne N-l+c-1a

Récapitulatif des tests à utiliser pour des variables quantitatives dans le cas de plus de 2 groupes (suite)

Tests	Rôle du test	Type de variable	Population	Conditions d'applications
Kruskall Wallis Analyse de variance à un facteur	Comparaison de plusieurs moyennes observées (plus de 2)	Quantitative	n échantillons indépendants étudiés selon un facteur	Aucune contrainte sur la forme des distributions Cas d'échantillons sans nœud Cas d'échantillons avec nœud

Tableau 13 (suite)

Formule	Table	Conclusions
$H = \frac{12}{N(N+1)} \sum_{j=1}^{n} \frac{R_j^2}{N_j} - 3(N+1)$		
1er cas : on dispose de 3 échantillons (n = 3) de petites tailles avec tous les $N_j \leq 5$, dans ce cas on utilise la table de Kruskall Wallis.	Kruskall Wallis	si $H_c \geq H_T$ alors on rejette H_0 au seuil α
2e cas : on dispose de 3 échantillons (n = 3) de grandes tailles avec tous les $N_j > 5$, ou de plus de 4 échantillons (n > 4) de grandes tailles avec tous les $N_j \geq 4$, dans ces cas on utilise la table du χ^2 à n-1 degrés de liberté, car dans ces cas la variable H suit une loi du χ^2 à n-1 ddl.	χ^2	si $H_c \geq \chi^2_T$ alors on rejette H_0 au seuil α avec ddl = n - 1
$H^* = \frac{\frac{12}{N(N+1)} \sum_{j=1}^{n} \frac{R_j^{*2}}{N_j} - 3(N+1)}{1 - \frac{\sum_{i=1}^{e} (d_i^3 - d_i)}{N^3 - N}}$		
3e cas : on dispose de 3 échantillons (n = 3) de petites tailles avec tous les Nj ≤ 5, dans ce cas on utilise la table de Kruskall Wallis avec H*.	Kruskall Wallis	si $H_c^* \geq H_T$ alors on rejette H_0 au seuil α
4e cas : si le nombre de termes égaux des différents nœuds de l'échantillon global excède 25 % de l'effectif de l'échantillon global, on utilise la table du χ^2 avec H*, avec ddl = n-1	χ^2	si $H_c^* \geq \chi^2_T$ alors on rejette H_0 au seuil α avec ddl = n - 1
5e cas : si le nombre de termes égaux des différents nœuds de l'échantillon global est inférieur à 25 % de l'effectif de l'échantillon global, on utilise la table du χ^2 avec H, avec ddl = n-1 Le pourcentage de termes égaux des différents nœuds de l'échantillon global est : $\frac{\text{nombre de termes avec nœuds}}{N}$	χ^2	si $H_c \geq \chi^2_T$ alors on rejette H_0 au seuil α avec ddl = n - 1

CHAPITRE IX

Exercices des chapitres VII et VIII

1. **Exercices corrigés** *259*

2. **Exercices non corrigés** *264*

3. **Solutions** *268*

CHAPITRE IX

Exercices des chapitres VII et VIII

1. Exercices corrigés (corrections pages 268-284)

• *Exercice 1 :*

Un psychologue soumet deux groupes de 50 sujets à deux tests A et B et note, pour chaque sujet, le temps de réponse en fin d'expérience ; il obtient les résultats suivants :

	Groupe A	Groupe B
Effectif	50	50
Moyenne expérimentale	200 secondes	218 secondes
Variance expérimentale	2000	3000

1. Les deux tests demandent-ils le même temps de réponse ?
2. Qu'en est-il si on accepte un risque de première espèce de 0,10 ?
3. Discuter les résultats, le cas échéant, en fonction du mode de constitution des deux groupes.

• *Exercice 2 :*

En psychologie expérimentale, on étudie l'effet de l'apprentissage chez des souris. Chez 40 souris soumises au test du labyrinthe, on dénombre les erreurs commises par chacune d'elles. Puis on réitère l'expérience chez ces souris, une demi-heure plus tard, dans les mêmes conditions, et on dénombre à nouveau les erreurs commises par chacune d'elles.
On souhaite déterminer si les souris font moins d'erreurs au cours de la

seconde expérience. On étudie donc la différence $d_i = x_i - y_i$ entre le nombre d'erreurs à la première et à la seconde expérience, où x_i désigne le nombre d'erreurs au premier test, et y_i le nombre d'erreurs au second test. Les résultats obtenus sont les suivants :

$\Sigma x_i = 3920$ $\Sigma x_i^2 = 387\,640$ $\Sigma d_i = 354$
$\Sigma y_i = 3566$ $\Sigma y_i^2 = 323\,350$ $\Sigma d_i^2 = 4482$
$\Sigma x_i y_i = 353\,254$

1. Donner la variation moyenne d'erreurs observée et l'intervalle de confiance de la variation moyenne d'erreurs chez les souris.
2. Comparer le nombre d'erreurs à la première expérience et à la seconde. Préciser les hypothèses de validité.

• *Exercice 3 :*

Au cours d'une étude sur la psychologie de l'enfant, on étudie chez 10 enfants de 18 mois le temps de réaction en situation d'"isolement". On filme des mères jouant avec leur enfant, puis sortant du champ visuel de l'enfant jusqu'à ce que celui-ci réagisse et réclame la présence de la mère. On mesure alors pour chaque enfant le temps de réaction, c'est-à-dire le temps entre le moment où leur mère s'absente et le moment où l'enfant manifeste l'absence de la mère. Les résultats sont les suivants :

Enfant	1	2	3	4	5	6	7	8	9	10
Temps de réaction (secondes)	25	20	12	30	30	10	10	15	15	8

On reproduit la même expérience chez ces enfants, vingt minutes plus tard. Les résultats sont les suivants :

Enfant	1	2	3	4	5	6	7	8	9	10
Temps de réaction (secondes)	30	25	15	40	38	10	14	20	18	8

Les temps de réaction sont-ils les mêmes d'une expérience à l'autre ?

• *Exercice 4 :*

On observe 10 personnes dans un labyrinthe de grande taille à deux reprises, à vingt-quatre heures d'intervalle, afin de déterminer l'effet de la mémoire du parcours. Pour cela on note le nombre d'erreurs commises par chaque personne pour chacune des épreuves, afin de comparer le nombre d'erreurs d'une épreuve à l'autre. Nous supposons donc que le nombre d'erreurs est plus important lors de la première épreuve comparativement à la seconde. Les données observées sont les suivantes :

Personnes	1	2	3	4	5	6	7	8	9	10
Epreuve 1 (X)	659	984	397	574	447	479	676	761	647	577
Epreuve 2 (X')	452	587	460	787	351	277	234	516	577	513

Poser les hypothèses et effectuer le test au seuil $\alpha = 5\ \%$.

• *Exercice 5 :*

On observe 30 sujets séropositifs au Virus d'Immunodéficience Humaine, pour lesquels on mesure à un an d'intervalle (à t_0 puis t_1) un score de qualité de vie, à partir d'une échelle de qualité de vie spécifique de l'infection à VIH (selon cette échelle, plus le score est important, meilleure est la qualité de vie). Le score de qualité de vie semble diminuer au cours de l'infection.
Poser clairement les hypothèses et effectuer le test au seuil $\alpha = 5\ \%$.
Nous avons noté les différences entre les scores d'une année à l'autre, noté Δ_k (score au temps t_0 moins score au temps t_1) :

Sujet k	Δk
1	- 2
2	0
3	0
4	1
5	0
6	0
7	4
8	4
9	1
10	1
11	5
12	3
13	5
14	3
15	- 1
16	1
17	- 1
18	5
19	8
20	2
21	2
22	2
23	- 3
24	- 2
25	1
26	4
27	8
28	2
29	3
30	- 1

•Exercice 6 :

Quatre laboratoires de recherche en psychologie sociale ont calculé pour huit de leurs chercheurs le nombre d'articles publiés les dix dernières années, en affectant un coefficient selon la renommée des revues de référence. Les scores de publication selon les laboratoires sont les suivants :

Labo 1	Labo 2	Labo 3	Labo 4
38,7	39,2	34	34
41,5	39,3	35	34,8
43,8	39,7	39	34,8
44,5	41,4	40	35,4
45,5	41,8	43	37,2
46,0	42,9	43	37,8
47,7	43,3	44	41,2
58,0	45,8	45	42,8

Poser les hypothèses et tester ces hypothèses au seuil $\alpha = 5\ \%$.

• Exercice 7 :

Une échelle de mesure de satisfaction a permis de calculer le score de satisfaction de trois groupes de sujets comparables (toutes choses égales par ailleurs) concernant trois produits. Le premier échantillon est composé de 100 sujets, le deuxième et le troisième échantillon sont composés de 50 sujets.
Les résultats sont présentés ci-dessous :

	effectif n	moyenne m	variance s^2
Produit 1	100	11,8	47
Produit 2	50	13,2	70
Produit 3	50	16,3	50

1. En terme de satisfaction, y a-t-il une différence globale entre ces trois produits ?
2. En terme de satisfaction, y a-t-il une différence entre le produit 1 et le produit 2 ?

• Exercice 8 :

Reprenons l'exemple présenté précédemment dans le paragraphe 2.2. du chapitre VIII (analyse de variance à deux facteurs croisés) avec des données différentes issues d'un autre tirage aléatoire.
On étudie le score de qualité de vie de 15 personnes atteintes d'un cancer du

foie, avant, pendant, et après une période de traitement par chimiothérapie. (Plus le score est élevé, meilleure est la qualité de vie). On cherche à déterminer si les scores de qualité de vie sont les mêmes selon ces trois périodes de temps, ou si au contraire ils diffèrent significativement dans l'ensemble. On cherche également à déterminer si les scores de qualité de vie sont les mêmes selon les personnes, ou si au contraire ils diffèrent significativement dans l'ensemble.
Les données sont les suivantes :

Personne	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Période															
Avant traitement	75	76	74	73	76	76	72	73	75	73	75	75	73	76	74
Pendant traitement	70	72	69	70	73	75	71	71	74	70	75	70	70	73	69
Après traitement	69	70	68	65	72	73	70	70	72	69	74	67	58	72	69

• *Exercice 9 :*

Les notes obtenues à un test de raisonnement, chez 10 élèves, à l'âge de 11 ans, à l'âge de 13 ans, puis à l'âge de 15 ans, sont présentées dans le tableau suivant :

Personne	1	2	3	4	5	6	7	8	9	10
Age										
Test à 11 ans	21	24	13	18	16	20	23	17	15	19
Test à 13 ans	30	32	25	28	27	31	32	28	24	30
Test à 15 ans	40	42	38	44	43	44	42	40	41	43

On cherche à déterminer si les notes observées sont les mêmes selon l'âge, ou si au contraire elles diffèrent significativement dans l'ensemble.
On cherche également à déterminer si ces notes sont les mêmes selon les élèves, ou si au contraire elles diffèrent significativement dans l'ensemble.

• *Exercice 10 :*

On étudie la suggestion hypnotique sur la fonction respiratoire, chez seize sujets répartis en deux groupes de huit sujets de façon aléatoire. Huit sujets sont destinés à être des sujets contrôle et les huit autres sont destinés à être hypnotisés.
On remarque qu'au début de l'expérience les mesures semblent faire apparaître une ventilation respiratoire plus importante chez les sujets à hypnotiser que chez les sujets contrôle. Une explication plausible serait un effet d'anticipation à être hypnotisé chez les sujets sachant faire partie du groupe traité. L'anticipation à l'hypnose provoque-t-elle une ventilation respiratoire particulièrement différente chez les sujets observés ? Tester cette hypothèse au seuil $\alpha = 5\ \%$.

Les données observées (mesures de la ventilation respiratoire) sont les suivantes :

échantillon contrôle (X)	3,99	4,19	4,21	4,54	4,64	4,69	4,84	5,48
échantillon hypnotisé (X')	4,36	4,67	4,78	5,08	5,16	5,20	5,52	5,74

2. Exercices non corrigés

• Exercice 11 :

On souhaiterait savoir si la notation d'une épreuve d'examen de psychologie clinique dépend ou non du correcteur. On recueille alors les notes proposées par dix correcteurs tirés au sort dans une population de correcteurs, chacun des correcteurs effectuant la notation d'une même copie deux fois (au cours d'une longue période de correction d'un grand nombre de copies). Les notations de chacun des correcteurs sont présentées dans le tableau suivant.

	Notes proposées par les dix correcteurs									
1re correction	5	7	8	9	7	12	10	13	2	8
2e correction	7	10	4	6	3	10	3	11	5	8

1. La notation est-elle dépendante du correcteur ?

2. Les correcteurs 8 et 9 doivent-ils être considérés comme atypiques ?

3. Dans quel intervalle situez-vous la vraie note ?

• Exercice 12 :

Dans une recherche en psychologie cognitive, on étudie le quotient intellectuel d'un groupe de 32 sujets au temps t_0. Puis on mesure à nouveau le QI de ces sujets au temps t_1.

On souhaite déterminer si les QI observés ont augmenté d'une période à l'autre. On étudie donc la différence $d_i = y_i - x_i$ entre les QI au temps t_1 et au temps t_0, où x_i désigne le QI au temps t_0, et y_i le QI au temps t_1. Les résultats obtenus sont les suivants :

$\Sigma x_i = 3324$ $\quad \Sigma x_i^2 = 356\,782$ $\quad \Sigma d_i = 252$

$\Sigma y_i = 3576$ $\quad \Sigma y_i^2 = 413\,206$ $\quad \Sigma d_i^2 = 2466$

$\Sigma x_i y_i = 383\,761$

1. Donner la variation moyenne du QI observée et l'intervalle de confiance de cette variation moyenne.

2. Comparer les QI au temps t_0 et au temps t_1. Préciser les hypothèses de validité.

• *Exercice 13 :*

Au cours d'une étude en psychologie différentielle, on étudie l'efficacité de deux modèles de psychologie de la vente. On souhaite déterminer quel est le modèle le plus efficace. On observe alors les recettes de quinze vendeurs de bateaux et de matériaux marins, représentants de commerce itinérants dans le sud de la France, selon ces deux modèles de vente. Pour chaque vendeur on dénombre les produits vendus et on note la recette correspondante, sur un mois, selon le premier modèle de vente ; puis le mois suivant selon le deuxième modèle de vente.

Les observations sont les suivantes :

Vendeurs	**Modèle de vente n°1**		**Modèle de vente n°2**	
	Nombre de produits vendus	Recette en Francs	Nombre de produits vendus	Recette en Francs
1	40	500 000	35	430 000
2	43	700 000	36	590 000
3	25	900 000	20	350 000
4	36	600 000	28	500 000
5	20	800 000	20	600 000
6	23	700 000	23	495 000
7	40	390 000	35	92 000
8	38	500 000	29	250 000
9	35	630 000	22	390 000
10	29	590 000	36	500 000
11	22	750 000	20	390 000
12	28	780 000	28	400 000
13	32	495 000	36	630 000
14	34	92 000	20	500 000
15	36	250 000	23	350 000

1. En terme de nombre de produits vendus quel est le modèle le plus efficace ?
2. En terme de recette quel est le modèle le plus efficace ?

• *Exercice 14 :*

Les résultats d'un test psychométrique mesurant le déficit cognitif, administrés à trois groupes de sujets épileptiques traités différemment, sont les suivants :

	Traitement A	Traitement B	Traitement C
Effectif (n)	100	80	70
Scores moyens (m)	18	22	26
Variance (s^2)	72	60	81

1. Y a-t-il une différence globale entre les scores moyens de déficit cognitif selon le traitement ?

2. Y a-t-il une différence entre les scores moyens selon les traitements A et B ?

• Exercice 15 :

On mesure le score d'anxiété chez des sujets "sains" habitant cinq villes différentes. Pour chacune des villes, le tableau suivant indique le nombre de sujets (n), la moyenne (m) et la variance (s^2) du score d'anxiété :

	Villes				
	V_1	V_2	V_3	V_4	V_5
n	50	50	50	50	40
m	3,5	3,7	3,5	3,6	3,8
s^2	0,42	0,34	0,37	0,40	0,39

1. Y a-t-il, en moyenne, une différence d'anxiété entre les habitants de ces cinq villes ?

2. Peut-on comparer les moyennes des villes V_1 et V_2 ?

• Exercice 16 :

Des échelles de mesure de stress, spécifiques au personnel soignant, ont permis de mesurer le score de stress des infirmières pratiquant dans des services de cancérologie de trois hôpitaux différents. Les résultats sont indiqués dans le tableau suivant :

	Hôpital 1	Hôpital 2	Hôpital 3
Nombre d'infirmières	38	50	32
moyenne	5,2	4,8	5,9
variance	2,0	2,8	2,3

On suppose que les variances vraies des scores de stress sont les mêmes dans les trois hôpitaux.

1. Donnez une estimation de cette variance.

2. Calculez l'intervalle de confiance de la moyenne des scores dans l'hôpital 3.

3. Les moyennes des scores de stress diffèrent-elles entre les hôpitaux ?

• *Exercice 17 :*

Dans une étude de psychologie du travail, on cherche à déterminer l'influence de trois programmes de réhabilitation vers le réemploi, chez 10 000 chômeurs à la suite d'un licenciement économique.

A cet effet, on prélève au hasard 20 sujets de chaque programme pour lesquels on relève ensuite le nombre de mois chômés depuis leur licenciement. Au total, 60 sujets sont donc inclus dans l'étude.

Ces chômeurs sont tous issus du même secteur d'activité initial et ont intégré le programme d'aide à la réhabilitation vers le réemploi dès leur licenciement. Les données résultant des tirages aléatoires sont les suivantes :

	Programmes de réhabilitation vers le réemploi			
	I	II	III	
Nombre de mois chômés par personne à la suite du licenciement	1	8	1	
	3	7	3	
	5	5	4	
	4	6	2	
	2	9	6	
	6	7	3	
	7	2	5	
	8	7	1	
	4	9	7	
	5	8	5	
	4	7	1	
	2	9	3	
	6	5	5	
	7	7	2	
	8	9	6	
	4	7	7	
	1	2	8	
	3	7	4	
	5	9	1	
	6	9	3	
	7	7	5	
Somme des x (T_i)	98	146	82	
Somme des x^2	550	1100	414	$\Sigma x^2 = 2064$

1. Peut-on dire que le nombre de mois chômés diffère, de façon significative, selon les programmes de réhabilitation vers le réemploi ?

2. Est-il nécessaire d'étudier la différence des nombres de mois chômés selon les programmes de réhabilitation vers le réemploi de façon détaillée (deux à deux) ? Dire pourquoi, et si oui, utiliser le test approprié.

3. S'il existe (ou s'il existait) une différence entre les programmes, peut-on (ou pourrait-on) attribuer cette différence à une meilleure efficacité à la réhabilitation de certains programmes ?

• *Exercice 18 :*

Dans une étude en psychologie différentielle, on cherche à évaluer la variation des temps de réponse à quatre auto-questionnaires, chez dix étudiants de psychologie d'une même université.
Les temps de réponse aux quatre questionnaires sont les suivants :

	Questionnaires			
	1	2	3	4
	15	15	14	17
	17	20	18	19
	16	22	15	15
Temps	14	14	15	20
de réponse	18	24	18	17
en minutes	15	14	15	16
des 10 sujets	15	12	20	14
	18	17	22	16
	16	16	14	15
	19	18	24	20

1. Peut-on dire que le temps de réponse diffère, de façon significative, selon le questionnaire et selon les sujets ?
2. Est-il nécessaire d'étudier la différence des temps de réponse selon les questionnaires de façon détaillée (deux à deux) ? Si oui, utiliser le test approprié.

3. Solutions

• *Correction de l'exercice 1 :*

1. Les hypothèses à tester pour répondre à la première question sont les suivantes :

H_0 : $\mu_A = \mu_B = \mu$: les moyennes vraies (ou théoriques) des temps de réponse des deux populations de référence des groupes observés sont équivalentes ;

contre

H_1 : $\mu_A \neq \mu_B$: les moyennes vraies (ou théoriques) des temps de réponse des deux populations de référence diffèrent ;

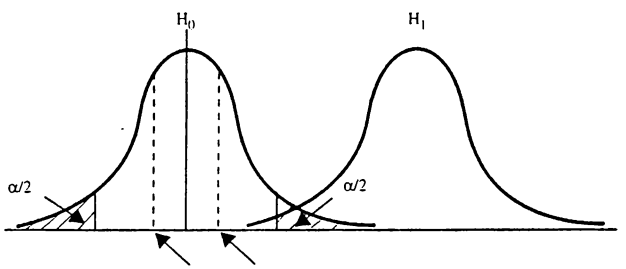
Pour tester ces hypothèses on effectue un test ε de comparaison de deux moyennes observées, les conditions d'application étant réalisées :
$n_A = n_B = 50$ (donc > 30).
On calcule alors le paramètre ε_c

$$\varepsilon_c = \frac{|m_A - m_B|}{\sqrt{\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}}}$$

$$\varepsilon_c = \frac{|200 - 218|}{\sqrt{\frac{2000}{50} + \frac{3000}{50}}} = 1{,}8$$

Au risque $\alpha = 5\ \%$, $\varepsilon_T = 1{,}96$ lu dans la table de la loi normale. On a $\varepsilon_c < \varepsilon_T$, donc on ne rejette pas l'hypothèse nulle. On n'a donc pas mis en évidence de différence significative entre les deux temps moyens de réponse au deux tests A et B.

2. Au risque $\alpha = 10\ \%$, $\varepsilon_T = 1{,}645$ lu dans la table de la loi normale. On a $\varepsilon_c > \varepsilon_T$, donc on rejette l'hypothèse nulle. On conclut donc à une différence significative entre les deux temps moyens de réponse aux deux tests A et B. Quand on augmente le seuil du risque de première espèce α, on finit parfois par conclure à une différence significative.



Augmentation du seuil du risque α

3. Au seuil $\alpha = 10\ \%$, on conclut à une différence significative, mais pas à un effet de causalité entre l'augmentation du temps de réponse et le test car les groupes n'ont pas été constitués par randomisation (ou tirage au sort). La différence observée est peut-être liée à la constitution du groupe (un groupe pouvant être plus lent que l'autre). Ainsi, un tirage au sort pour la constitution des deux groupes aurait été nécessaire pour conclure à l'existence d'un lien de causalité entre le type de test et le temps de réponse.

• Correction de l'exercice 2 :

1. $\Sigma d_i = 354 \qquad \Sigma d_i^2 = 4482$

Variation moyenne = la moyenne des variations

$$= m_d = \frac{\sum d_i}{n} = \frac{354}{40} = 8{,}85$$

Variance estimée

$$= s_d^2 = \frac{1}{n-1}\left[\sum d_i^2 - \frac{\left(\sum d_i\right)^2}{n}\right]$$

$$= \frac{1}{39}\left[4482 - \frac{(354)^2}{40}\right] = 34{,}59$$

$$IC_{(95\%)} = \left[m_d - \varepsilon_{5\%}\sqrt{\frac{s_d^2}{n}} \;;\; m_d + \varepsilon_{5\%}\sqrt{\frac{s_d^2}{n}}\right]$$

$$IC_{(95\%)} = \left[8{,}85 - 1{,}96\sqrt{\frac{34{,}59}{40}} \;;\; 8{,}85 + 1{,}96\sqrt{\frac{34{,}59}{40}}\right]$$

$$IC_{(95\%)} = [7{,}03 \;;\; 10{,}67]$$

les conditions d'application étant vérifiées, car $n \geq 30$.

2. Pour répondre à la deuxième question les hypothèses à tester sont :

H_0 : Le nombre moyen d'erreurs à la première expérience ne diffère pas significativement de celui correspondant à la seconde expérience. On n'observe pas d'effet d'apprentissage chez les 40 souris. Soit $\mu_{E1} = \mu_{E2}$.

contre

H_1 : Le nombre moyen d'erreurs à la première expérience diffère significativement de celui correspondant à la seconde expérience. On observe un effet de l'apprentissage chez les 40 souris. Soit $\mu_{E1} \neq \mu_{E2}$.

S'agissant de mesures effectuées chez les mêmes souris en deux temps, les deux séries de données sont appariées. Un test de comparaison de deux moyennes observées dans le cas de séries appariées sera donc effectué. Les conditions d'application étant vérifiées $n \geq 30$, un test ε (apparié) peut donc être effectué pour tester les hypothèses :

$$\begin{cases} H_0 : \mu_d = 0 \\ H_1 : \mu_d \neq 0 \end{cases}$$
contre

avec μ_d correspondant à la moyenne des différences du nombre d'erreurs d'une expérience à l'autre.

Le paramètre à calculer est alors :

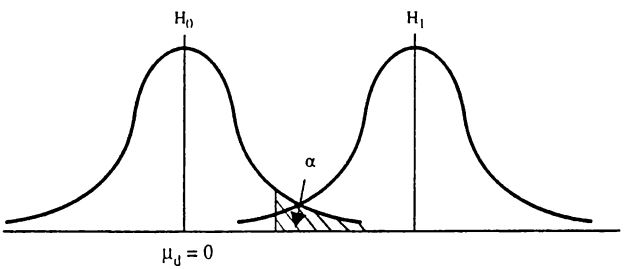
$$\varepsilon_c = \frac{m_d - 0}{\sqrt{\frac{s_d^2}{n}}}$$

$$\varepsilon_c = \frac{8,85}{\sqrt{\frac{34,59}{40}}} = 9,517$$

Au seuil $\alpha = 5\ \%$, $\varepsilon_T = 1,96$, lu dans la table de la loi normale. Ainsi, $\varepsilon_c > \varepsilon_T$, on conclut donc au rejet de H_0 ($p < 0,000\ 000\ 001$). La moyenne d'erreurs commises par les souris au premier test du labyrinthe est significativement différente de celle correspondant au deuxième test.
Dans le cas d'un test unilatéral à droite :

$$\begin{cases} H_0 : \mu_d = 0 \\ H_1 : \mu_d > 0 \end{cases}$$
contre

Au seuil $\alpha = 5\ \%$, $\varepsilon_T = 1,645$, lu dans la table de la loi normale.



Ainsi, on obtient encore $\varepsilon_c > \varepsilon_T$, on conclut donc au rejet de H_0 ($p < 0,000\ 000\ 001$). La moyenne d'erreurs commises par les souris au premier test du labyrinthe est significativement supérieure à celle correspondant au deuxième test.

• *Correction de l'exercice 3 :*

Afin de déterminer si les temps de réaction sont les mêmes d'une expérience à l'autre, les hypothèses à tester sont :

H_0 : Le temps moyen de réaction à la première expérience ne diffère pas significativement de celui correspondant à la seconde expérience. Soit $\mu_{E1} = \mu_{E2}$

contre

H_1 : Le temps moyen de réaction à la première expérience diffère significativement de celui correspondant à la seconde expérience. Soit $\mu_{E1} \neq \mu_{E2}$

S'agissant de mesures effectuées chez les mêmes enfants en deux temps, les deux séries de données sont appariées. Un test de comparaison de deux moyennes observées dans le cas de séries appariées sera donc effectué. Les conditions d'application n'étant pas vérifiées, $n < 30$, pour réaliser un test ε (apparié), et en supposant que la différence des temps de réaction est distribuée normalement, on teste les hypothèses par un test t (apparié).

Les hypothèses sont alors :

$$\begin{cases} H_0 : \mu_d = 0 \\ \text{contre} \\ H_1 : \mu_d \neq 0 \end{cases}$$

avec μ_d correspondant à la moyenne vraie des différences des temps de réaction.

Le paramètre à calculer est alors :

$$t_c = \frac{m_d - 0}{\sqrt{\frac{s_d^2}{n}}}$$

Il faut alors calculer m_d et s_d^2 à partir du tableau suivant :

Enfant	1	2	3	4	5	6	7	8	9	10	
Temps de réaction Première expérience	25	20	12	30	30	10	10	15	15	8	
Temps de réaction Seconde expérience	30	25	15	40	38	10	14	20	18	8	
Différences d_i (Exp.2 - Exp.1)	5	5	3	10	8	0	4	5	3	0	Σd_i 43
Différences au carré d_i^2	25	25	9	100	64	0	16	25	9	0	Σd_i^2 273

Moyenne des différences : $m_d = \frac{\sum d_i}{n} = \frac{43}{10} = 4{,}3$

Variance estimée des différences : $s_d^2 = \frac{1}{n-1}\left[\sum d_i^2 - \frac{\left(\sum d_i\right)^2}{n}\right]$

$$= \frac{1}{9}\left[273 - \frac{(43)^2}{10}\right] = 9{,}79$$

Le paramètre calculé est alors :

$$t_c = \frac{4,3}{\sqrt{\frac{9,79}{10}}} = 4,346$$

Au seuil $\alpha = 5\ \%$, $t_T = 2,262$, lu dans la table de t, avec n-1 = 9 degrés de liberté. Ainsi, $t_c > t_T$, on conclut donc au rejet de H_0, la moyenne des différences diffère significativement de 0. Alors, les moyennes observées diffèrent significativement, c'est-à-dire que la moyenne du temps de réaction des enfants à la première expérience est significativement différente de celle correspondant à la deuxième expérience.

• *Correction de l'exercice 4 :*

Les hypothèses à tester sont :

$\left\{ \begin{array}{l} H_0 : X \text{ et } X' \text{ sont de même loi} \\ H_1 : \text{loi de } X > \text{loi de } X' \text{ aléatoirement} \end{array} \right.$

Soit H_0 : le nombre d'erreurs est équivalent d'une épreuve à l'autre

contre H_1 : le nombre d'erreurs à la 1re épreuve est supérieur à celui de la 2e épreuve

La taille de l'échantillon est n = 10.

Il s'agit de deux échantillons couplés (ou appariés) et il s'agit donc de comparer les deux lois par un test de Wilcoxon dans le cas d'un petit échantillon sans nœud. Pour cela il faut calculer les différences Δ_k.

Personnes	1	2	3	4	5	6	7	8	9	10
Epreuve 1 (X)	659	984	397	574	447	479	676	761	647	577
Epreuve 2 (X')	452	587	460	787	351	277	234	516	577	513
Δ_k	207	397	-63	-213	96	202	442	245	70	64

$\lvert\Delta_k\rvert$ ordonnés de façon croissante et assignés de leur rang :										
$\lvert\Delta_k\rvert$	63	64	70	96	202	207	213	245	397	442
rang	-1	2	3	4	5	6	-7	8	9	10

On calcule la somme des rangs positifs et la somme des rangs négatifs :

$W_+ = 2 + 3 + 4 + 5 + 6 + 8 + 9 + 10 = 47$

et $W_- = -1 + (-7) = -8$

Notons W_C, la plus petite des deux valeurs W_+ ou W_- (C pour calculé).

La table de Wilcoxon permet de déterminer la limite supérieure accordée à W_C pour conclure au rejet de H_0. Ici, il s'agit alors de W_- que l'on compare à la valeur donnée par la table de Wilcoxon, W_T (T pour tabulé), pour $\alpha = 5\%$.
La limite supérieure accordée à W_C pour le rejet de H_0 est $W_T = 8$, or $W_C = 8$.
Autrement dit :
si pour $\alpha = 5\ \%$, $W_C > W_T$ alors on conclut au non rejet de H_0.
si pour $\alpha = 5\ \%$, $W_C \leq W_T$ alors on conclut au rejet de H_0.
De même, si pour $\alpha = 1\ \%$, $W_C \leq W_T$ alors on conclut encore au rejet de H_0, et $\alpha = 1\ \%$ fixe le degré de signification de la différence entre les deux séries observées.
Donc au seuil $\alpha = 5\ \%$ on conclut au rejet de H_0, c'est-à-dire à une différence significative entre les deux séries de données.
En revanche, pour $\alpha = 1\ \%$ on ne conclut plus à une différence significative, puisqu'on lit $W_T = 3$. Alors le degré de signification de cette différence entre les deux séries de données appariées est $p < 0{,}05$.
On conclut que le nombre d'erreurs diminue d'une épreuve à l'autre. Il semble donc que le parcours du labyrinthe soit partiellement mémorisé à 24 heures d'intervalle, à moins que des facteurs de confusion ne soient pas contrôlés.

• *Correction de l'exercice 5 :*

Les hypothèses à tester sont :

$\left\{ \begin{array}{l} H_0 : X \text{ et } X' \text{ sont de même loi} \\ H_1 : \text{loi de } X > \text{loi de } X', \text{ aléatoirement} \end{array} \right.$

Soit H_0 : les scores de qualité de vie sont équivalents d'une année sur l'autre

contre H_1 : les scores de qualité de vie sont supérieurs à t_0 comparativement à t_1

La taille de l'échantillon est $n = 26$ une fois les nœuds éliminés dans le cas où $\Delta_k = 0$ (voir p. 212)
Il s'agit de deux échantillons couplés (ou appariés) et il s'agit donc de comparer les deux lois par un test de Wilcoxon dans le cas d'un grand échantillon avec nœud. Pour cela il faut déterminer les rangs des $|\Delta_k|$.

Sujet k	Δ_k	rangs naturels	rangs
1	- 2	9	-11,5
2	0	éliminé	éliminé
3	0	éliminé	éliminé
4	1	1	4,5
5	0	éliminé	éliminé
6	0	éliminé	éliminé
7	4	19	20
8	4	20	20
9	1	2	4,5
10	1	3	4,5
11	5	22	23
12	3	15	16,5
13	5	23	23
14	3	16	16,5
15	- 1	4	- 4,5
16	1	5	4,5
17	- 1	6	- 4,5
18	5	24	23
19	8	25	25,5
20	2	10	11,5
21	2	11	11,5
22	2	12	11,5
23	- 3	17	- 16,5
24	- 2	13	- 11,5
25	1	7	4,5
26	4	21	20
27	8	26	25,5
28	2	14	11,5
29	3	18	16,5
30	- 1	8	- 4,5

Le calcul des rangs a été effectué de la façon suivante, pour chaque différence $|\Delta_k|$:

$$|\Delta_k| = 1 \quad \Rightarrow \quad \frac{1+2+3+4+5+6+7+8}{8} = 4,5$$

$$|\Delta_k| = 2 \quad \Rightarrow \quad \frac{9+10+11+12+13+14}{6} = 11,5$$

$$|\Delta_k| = 3 \quad \Rightarrow \quad \frac{15+16+17+18}{4} = 16,5$$

$$|\Delta_k| = 4 \quad \Rightarrow \quad \frac{19+20+21}{3} = 20$$

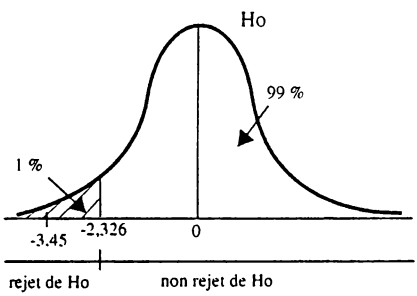
$|\Delta_k| = 5 \quad \Rightarrow \quad \frac{22+23+24}{3} = 23$

$|\Delta_k| = 8 \quad \Rightarrow \quad \frac{25+26}{2} = 25,5$

On calcule $W_- = 2\,|-11,5| + 3\,|-4,5| + 16,5 = 53$
qui correspond à un test unilatéral à gauche
puis on calcule

$$\varepsilon_c = \frac{W_- - \frac{n(n+1)}{4}}{\sqrt{\frac{n(n+1)(2n+1)}{24}}} = \frac{53 - \frac{26\times 27}{4}}{\sqrt{\frac{26\times 27\times 53}{24}}} = -3,11$$

Nous sommes ici dans le cas d'un test unilatéral à gauche, donc la table de la loi normale centrée réduite N(0, 1) permet de déterminer le paramètre tabulé ε_T, à comparer à ε_C, le paramètre calculé précédemment.



Pour $\alpha = 1\ \%$, $\varepsilon_T = -2,326$, donc on rejette encore l'hypothèse H_0 car $\varepsilon_T > \varepsilon_C$. Ainsi on rejette l'hypothèse H_0 et on conclut que les scores de qualité de vie diminuent d'une année sur l'autre chez les sujets observés ($p < 0,01$).

• *Correction de l'exercice 6 :*

Les hypothèses à tester sont :

soit — H_0 : il n'y a pas de différence entre les scores de publication des quatre laboratoires

contre — H_1 : il y a une différence entre les scores de publication des quatre laboratoires

Un test de Kruskall Wallis dans le cas d'échantillons avec nœud permettra de tester ces hypothèses.

Labo 1 score	Labo 2 score	Labo 3 score	Labo 4 score
38,7	39,2	*34*	*34*
41,5	39,3	35	*34,8*
43,8	39,7	39	*34,8*
44,5	41,4	40	35,4
45,5	41,8	*43*	37,2
46,0	42,9	*43*	37,8
47,7	43,3	44	41,2
58,0	45,8	45	42,8

Six termes (termes en italique) interviennent dans les nœuds, soit moins de 20 % (6/32), donc on calcule H.
Pour cela on calcule la somme des rangs de chaque échantillon :

$R_1 = 197 \quad R_2 = 142 \quad R_3 = 125 \quad R_4 = 64$

Alors $$H = \frac{12}{N(N+1)} \sum_{j=1}^{n} \frac{R_j^2}{N_j} - 3(N+1)$$

soit $$H = \frac{12}{32 \times 33}\left(\frac{(197)^2}{8} + \frac{(142)^2}{8} + \frac{(125)^2}{8} + \frac{(64)^2}{8}\right) - 3 \times 33$$

$$H = 12{,}78$$

On lit la valeur du paramètre tabulé dans la table du χ^2 pour $\alpha = 5\ \%$, pour ddl = n-1 = 3, soit $H_T = 7{,}81$
Ainsi $H_C > H_T$ donc on rejette l'hypothèse H_0 et on conclut que la différence entre les quatre laboratoires en terme de publication est significative ($p < 0{,}01$). Par la suite il serait intéressant de rechercher quel est le laboratoire qui publie le plus.

• *Correction de l'exercice 7 :*

1. Les hypothèses correspondant à la question posée sont :

H_0 : Les moyennes des scores de satisfaction aux 3 produits ne diffèrent pas significativement.
H_1 : Les moyennes des scores de satisfaction aux 3 produits diffèrent significativement dans l'ensemble.

En supposant que le score de satisfaction soit distribué normalement et de même variance pour les trois échantillons, alors la comparaison de ces trois moyennes sera effectuée par une analyse de variance à un facteur.

Ne disposant pas des données individuelles on déduit le calcul de $\Sigma T_i^2 / n_i$ par :

$$\Sigma T_i^2 / n_i = \Sigma n_i m_i^2$$

car pour chaque produit i : $m_i = \Sigma x / n_i$, et $T_i = \Sigma x$

La variance résiduelle est obtenue directement en calculant :

$$s^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 + (n_3 - 1)s_3^2}{(n_1 - 1) + (n_2 - 1) + (n_3 - 1)}$$

$$s^2 = \frac{99 \times 47 + 49 \times 70 + 49 \times 50}{99 + 49 + 49} = 53,5$$

On a donc : $F_c = \frac{337,7}{53,5} = 6,31$

Le tableau d'analyse de variance permettant de comparer globalement les trois produits est le suivant.

Tableau d'analyse de variance :

Origine	Somme des carrés des écarts	d.d.l.	Variance	Fc
Entre produits	$\Sigma T_i^2 / n_i - T_G^2 / N$ $= \Sigma n_i m_i^2 - T_G^2 / N$ $= 35920,5 - 2655^2/200$ $= 675,4$	3-1=2	675,4/2 $= 337,7$	$\frac{337,7}{53.5} = 6,31$
Résiduelle		200-3=197	53,5	
Pour $\alpha = 1\ \%$ $F^2_{197} = 4,61$ donc $F_c = 6,31 > F^2_{197}$. Conclusion : en terme de satisfaction des sujets, les trois produits diffèrent significativement ($p < 0,01$).				

2. Pour comparer les moyennes théoriques des scores de satisfaction des sujets au produit 1 et au produit 2 (à partir des moyennes observées $m_1 = 11,8$ et $m_2 = 13,2$), on réalise un test de comparaison de deux moyennes (test t, la variance commune aux deux échantillons correspondant à la variance résiduelle). La question que l'on se pose est de savoir si les deux échantillons sont issus d'une même population de référence de moyenne commune μ, ou si au contraire ils sont issus de deux populations différentes de moyennes théoriques différentes. Les hypothèses à tester sont alors les suivantes :

$$\begin{cases} H_0 : \mu_1 = \mu_2 = \mu \\ H_1 : \mu_1 \neq \mu_2 \end{cases}$$

C'est-à-dire :

H_0 : La moyenne des scores de satisfaction du produit 1 est équivalente à celle du produit 2.

H_1 : La moyenne des scores de satisfaction du produit 1 diffère significativement de celle du produit 2.

On calcule alors :

$$t_c = \frac{|m_1 - m_2|}{\sqrt{\frac{s^2}{n_1} + \frac{s^2}{n_2}}}$$

$$t_c = \frac{|11,8 - 13,2|}{\sqrt{\frac{53,5}{50} + \frac{53,5}{100}}} = \frac{1,4}{1,267} = 1,10$$

Le nombre de degrés de liberté du t "tabulé" est celui de la variance résiduelle, c'est-à-dire 197, soit $t_T = 1,96$; alors $t_c < t_T$.

Au risque $\alpha = 5\ \%$, on ne met donc pas en évidence de différence entre ces deux moyennes, ainsi on ne met pas en évidence de différence significative entre les produits 1 et 2, en terme de satisfaction des sujets.

• *Correction de l'exercice 8 :*

Dans le but de tester l'effet période, les hypothèses sont les suivantes :

- H_0 : les scores moyens de qualité de vie sont les mêmes selon ces trois périodes de temps,
- H_1 : les scores moyens de qualité de vie diffèrent significativement dans l'ensemble selon ces trois périodes de temps.

Dans le but de tester l'effet du facteur personne, les hypothèses sont les suivantes :

- H_0 : les scores moyens de qualité de vie sont les mêmes selon les personnes observées,
- H_1 : les scores moyens de qualité de vie diffèrent significativeme dans l'ensemble selon les personnes observées.

En supposant que les conditions d'application soient vérifiées, alors on teste ces hypothèses par une analyse de variance à deux facteurs.

Données diminuées de 71 et calculs des paramètres entrant en jeu dans l'analyse de variance :

Personne	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	T_i	T_i^2	n_i
Période																		
Avant Tmt	4	5	3	2	5	5	1	2	4	2	4	4	2	5	3	51	2601	15
Pendant Tmt	-1	1	-2	-1	2	4	0	0	3	-1	4	-1	-1	2	-2	7	49	15
Après Tmt	-2	-1	-3	-6	1	2	-1	-1	1	-2	3	-4	-13	1	-2	-27	729	15
T_j	1	5	-2	-5	8	11	0	1	8	-1	11	-1	-12	8	-1	$T_G = 31$	$\Sigma T_i^2 = 3379$	
T_j^2	1	25	4	25	64	121	0	1	64	1	121	1	144	64	1	$\Sigma T_j^2 = 637$		
$\Sigma_j x^2$	21	27	22	41	30	45	2	5	26	9	41	33	174	30	17	$\Sigma x^2 = 523$		
n_j	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3			N=45

Tableau d'analyse de variance :

Origine	Somme des carrés des écarts	d.d.l.	Variance	Fc
Entre périodes	$\Sigma T_i^2 / n_i - T_G^2 / N$ = 3379/15 - 31²/45 = 203,91	3-1=2	203,91/2 = 101,96	$\frac{101,96}{3,81} = 26,74$
Entre personnes	$\Sigma T_j^2 / n_j - T_G^2 / N$ = 637/3 - 31²/45 = 190,98	15-1=14	190,98/14 = 13,64	$\frac{13,64}{3,81} = 3,58$
Résiduelle : Total - (Entre périodes + Entre personnes)	501,64 - (203,91 + 190,98) = 106,76	44-16=28	106,76/28 = 3,81	
Total	$\Sigma x^2 - T_G^2 / N$ = 523 - 31²/45 = 501,64	45-1=44		

Pour $\alpha = 1$ ‰ $F_{28}^{2} = 8,93$ donc $F_{c1} = 26,74 > F_{28}^{2}$.
Conclusion : le facteur période est significatif ($p < 0,001$).
Pour $\alpha = 1$ % $F_{28}^{14} = 2,75$ donc $F_{c2} = 3,58 > F_{28}^{14}$.
Conclusion : le facteur personne est significatif ($p < 0,01$).

Encore une fois, il existe aussi bien un facteur personne qu'un facteur période. On peut donc conclure à une différence significative entre les moyennes des scores de qualité de vie, dans l'ensemble, avant, pendant et après le traitement ($p < 0,001$). Cette différence est apparement dans le sens d'une diminution de la qualité de vie au cours du temps. De plus, on conclut à une différence significative entre les moyennes des scores de qualité de vie, dans l'ensemble, selon les personnes atteintes d'un cancer ($p < 0,01$).
Les scores moyens de qualité de vie étant différents significativement dans l'ensemble selon les périodes de traitement, une comparaison des scores moyens par période deux à deux permettra de déterminer plus en détail où se situe cette différence. Le test t (compte tenu des effectifs) de comparaison de deux moyennes observées sera alors utilisé.

	Période		
	Avant traitement	Pendant traitement	Après traitement
Moyenne	74,40	71,47	69,20

La variance commune aux trois périodes correspondant à la variance résiduelle, $s^2 = 3,81$.

Comparaison "Avant traitement" et "Pendant traitement" :

$$t_c = \frac{74,40 - 71,47}{\sqrt{\frac{3,81}{15} + \frac{3.81}{15}}} = \frac{2,93}{\sqrt{0,508}} = 4,11$$

La valeur de t "tabulé", lue dans la table de t pour le nombre de degrés de liberté de la variance résiduelle (d.d.l. = 28), au risque 5 %, étant 2,048 ; la valeur de t calculé est supérieure à la valeur de t "tabulé", soit $t_c > t_T$. Ainsi, la différence entre les deux moyennes, observées avant et pendant le traitement, est significative, avec un degré de signification de 1 ‰ ($p < 0,001$).

Comparaison "Pendant traitement" et "Après traitement" :

$$t_c = \frac{71,47 - 69,20}{\sqrt{\frac{3,81}{15} + \frac{3.81}{15}}} = \frac{2,27}{\sqrt{0,508}} = 3,18$$

On obtient $t_c > t_T$. Ainsi, la différence entre les deux moyennes, observées pendant et après le traitement, est significative, avec un degré de signification de 1 ‰ ($p < 0,001$).

Comparaison "Avant traitement" et "Après traitement" :

$$t_c = \frac{74,40 - 69,20}{\sqrt{\frac{3,81}{15} + \frac{3.81}{15}}} = \frac{5,2}{\sqrt{0,508}} = 7,30$$

On obtient à nouveau $t_c > t_T$. Ainsi, la différence entre les deux moyennes, observées avant et après le traitement, est significative, avec un degré de signification de 1 ‰ ($p < 0,001$).

Ainsi, les scores moyens de qualité de vie diffèrent significativement dans l'ensemble et deux à deux selon la période de traitement, dans le sens d'une diminution de la qualité de vie au cours du temps.

• *Correction de l'exercice 9 :*

Dans le but de tester l'effet âge, les hypothèses sont les suivantes :

H_0 : les moyennes des notes au test de raisonnement sont les mêmes selon le facteur âge,

H_1 : les moyennes des notes au test de raisonnement diffèrent significativement dans l'ensemble selon le facteur âge.

Dans le but de tester l'effet du facteur personne, les hypothèses sont les suivantes :

H_0 : les moyennes des notes au test de raisonnement sont les mêmes selon les élèves,

H_1 : les moyennes des notes au test de raisonnement diffèrent significativement dans l'ensemble selon les élèves.

En supposant que les conditions d'application soient vérifiées, alors on teste ces hypothèses par une analyse de variance à deux facteurs.

Données diminuées de 30 et calculs des paramètres entrant en jeu dans l'analyse de variance :

Personne	1	2	3	4	5	6	7	8	9	10	T_i	T_i^2	n_i
Age													
Test à 11 ans	-9	-6	-17	-12	-14	-10	-7	-13	-15	-11	-114	12996	10
Test à 13 ans	0	2	-5	-2	-3	1	2	-2	-6	0	-13	169	10
Test à 15 ans	10	12	8	14	13	14	12	10	11	13	117	13689	10
T_j	1	8	-14	0	-4	5	7	-5	-10	2	$T_G = -10$	$\Sigma T_i^2 = 26854$	
T_j^2	1	64	196	0	16	25	49	25	100	4	$\Sigma T_j^2 = 480$		
$\Sigma_i x^2$	181	184	378	344	374	297	197	273	382	290	$\Sigma x^2 = 2900$		
n_j	3	3	3	3	3	3	3	3	3	3			N=30

Analyse de variance : facteur âge et facteur personne

Origine	Somme des carrés des écarts	d.d.l.	Variance	Fc
Entre âges	$\Sigma T_i^2 / n_i - T_G^2 / N$ $= 26854/10 - 10^2/30 = 2682$	3-1=2	2682/2 $= 1341$	$\frac{1341}{3,22} = 416.46$
Entre personnes	$\Sigma T_j^2 / n_j - T_G^2 / N$ $= 480/3 - 10^2/30 = 156.67$	10-1=9	156,67/9 $= 17,4$	$\frac{17,4}{3.22} = 5,4$
Résiduelle : Total - (Entre périodes + Entre personnes)	$501.64 - (203,91 + 190,98)$ $= 57.9$	29-11=18	57,9/18 $= 3,22$	
Total	$\Sigma x^2 - T_G^2 / N$ $= 2900 - 10^2/30 = 2896.67$	30-1=29		

A 1 ‰ $F_{18}^2 = 10,39$ donc $F_{c1} = 416,46 > F_{18}^2$.

Conclusion : le facteur âge est significatif ($p < 0,001$)

A 1 % $F_{18}^9 = 3,6$ donc $F_{c2} = 5,4 > F_{18}^9$.

Conclusion : le facteur personne est significatif ($p < 0,01$)

Il existe aussi bien un facteur personne qu'un facteur période. On peut donc conclure à une différence significative entre les moyennes des notes au test

de raisonnement, dans l'ensemble, selon l'âge ($p < 0,001$). De plus, on conclut à une différence significative entre les moyennes des notes au test de raisonnement, dans l'ensemble, selon les élèves ($p < 0,01$).
Les moyennes des notes au test de raisonnement étant différentes significativement dans l'ensemble selon l'âge, une comparaison des moyennes des notes par âge deux à deux permettra de déterminer plus en détail où se situe cette différence. Le test t (compte tenu des effectifs) de comparaison de deux moyennes observées sera alors utilisé.

	Age		
	11 ans	13 ans	15 ans
Moyenne des notes	18,60	28,70	41,70

La variance commune aux trois âges correspondant à la variance résiduelle, $s^2 = 3,22$.

Comparaison "test à 11 ans" et "test à 13 ans" :

$$t_c = \frac{28,7 - 18,6}{\sqrt{\frac{3,22}{10} + \frac{3,22}{10}}} = \frac{10,1}{\sqrt{0,644}} = 12,58$$

La valeur de t "tabulé", lue dans la table de t pour le nombre de degrés de liberté de la variance résiduelle (d.d.l. = 18), au risque 5 %, étant 2,101 ; la valeur de t calculé est supérieure à la valeur de t "tabulé", soit $t_c > t_T$. Ainsi, la différence entre les deux moyennes des notes au test de raisonnement, à 11 ans et à 13 ans, est significative, avec un degré de signification de 1 ‰ ($p < 0,001$).

Comparaison "test à 13 ans" et "test à 15 ans" :

$$t_c = \frac{41,7 - 28,7}{\sqrt{\frac{3,22}{10} + \frac{3,22}{10}}} = \frac{13}{\sqrt{0,644}} = 16,2$$

On obtient $t_c > t_T$. Ainsi, la différence entre les deux moyennes des notes, à 11 ans et à 13 ans, est significative, avec un degré de signification de 1 ‰ ($p < 0,001$).

Comparaison "test à 11 ans" et "test à 15 ans" :

$$t_c = \frac{41,7 - 18,6}{\sqrt{\frac{3,22}{10} + \frac{3,22}{10}}} = \frac{23,1}{\sqrt{0,644}} = 28,78$$

Comme on pouvait s'y attendre, on obtient à nouveau $t_c > t_T$. Ainsi, la différence entre les deux moyennes des notes, à 11 ans et à 13 ans, est significative, avec un degré de signification de 1 ‰ ($p < 0,001$).

Ainsi, les moyennes des notes au test de raisonnement diffèrent significativement dans l'ensemble et deux à deux selon l'âge, dans le sens d'une augmentation de la note avec l'âge.

• *Correction de l'exercice 10 :*

Les hypothèses à tester, au seuil $\alpha = 5\ \%$, sont :

H_0 : X et X' sont de même loi

contre

H_1 : X et X' sont de lois différentes

La taille des échantillons est respectivement $n_1 = 8$ et $n_2 = 8$. On effectue alors un test de Mann Whitney dans le cas où $n_1 \leq 8$ et $n_2 \leq 8$.

L'échantillon global ordonné est :

	3,99	4,19	4,21	4,36	4,54	4,64	4,67	4,69
échantillon	1	1	1	2	1	1	2	1
rang	1	2	3	4	5	6	7	8
	4,78	4,84	5,08	5,16	5,20	5,48	5,52	5,74
échantillon	2	1	2	2	2	1	2	2
rang	9	10	11	12	13	14	15	16

$$R_1 = 1 + 2 + 3 + 5 + 6 + 8 + 10 + 14 = 49$$

$$U_1 = 49 - \frac{8 \times 9}{2} = 49 - 36 = 13$$

$$R_2 = 4 + 7 + 9 + 11 + 12 + 13 + 15 + 16 = 87$$

$$U_2 = 87 - 36 = 51$$

Alors $U_C = \min(U_1, U_2) = 13$

On lit alors dans la table de Mann et Whitney la valeur de U_T, au seuil $\alpha = 5\ \%$, à l'intersection de la colonne 8 et de la ligne 0 (puisque $n_1 = 8$ et $n_2 = 8$), soit $U_T = 13$.

On a $U_C = U_T$, dans ce cas on rejette l'hypothèse H_0, et on conclut à une différence significative entre les deux lois. C'est-à-dire, que la ventilation respiratoire est plus importante chez les sujets de l'échantillon susceptibles d'être sous hypnose que chez les sujets contrôle observés. Mais seule une étude plus approfondie sur davantage de sujets permettrait de conclure à un lien de causalité entre l'anticipation à l'hypnose et une augmentation de la ventilation respiratoire en règle générale.

CHAPITRE X

Tests de liaison entre deux variables quantitatives

1. Test d'indépendance entre deux variables quantitatives *287*
1.1. Adéquation entre le test et la question posée *287*
1.2. Définition des termes corrélation et régression *288*

2. La régression *289*
2.1. Indépendance entre deux variables quantitatives *289*
2.2. Ajuster une droite au sens des moindres carrés *293*
2.3. Analyse de l'équation de régression *293*
2.3.1. La pente de la droite de régression *293*
2.3.2. Test sur la pente p de la droite de régression *296*
2.3.3. Linéarité de la droite de régression *297*
2.4. Intervalles de confiance *298*
2.4.1. Intervalle de confiance de la droite de régression *298*
2.4.2. Prévision pour un nouveau point d'abcisse x_i *299*
2.5. Analyse des résidus *300*
2.5.1. Vérification des conditions d'application *300*
2.5.2. Test des séquences sur les résidus *302*
2.5.3. Distribution des résidus *303*

3. La corrélation *303*
3.1. Le coefficient de corrélation de Bravais-Pearson *304*
3.1.1. Rôle du test de Bravais-Pearson *304*
3.1.2. Procédure *304*
3.1.3. Lecture de la table du coefficient de corrélation *305*
3.1.4. Rôle et variations du coefficient de corrélation *306*
3.2. Test et intervalle de confiance *307*
3.3. La corrélation sur les rangs *308*
3.3.1. Rôle du test de corrélation sur les rangs *308*
3.3.2. Procédure *308*

4. Utilisation pertinente de la régression et/ou corrélation *309*

5. Exercices *311*
5.1. Exercices corrigés *311*
5.2. Exercices non corrigés *312*
5.3. Solutions *313*

CHAPITRE X

Tests de liaison entre deux variables quantitatives

1. Test d'indépendance entre deux variables quantitatives

Nous allons, dans ce chapitre, étudier l'association entre deux variables X et Y. Il s'agira, dans tous les cas, de variables quantitatives comme nous les avons définies au chapitre I. Jusqu'à présent, nous avions comparé les différentes réalisations d'une seule et même variable quantitative, que ce soit pour une comparaison de moyennes ou de variance. Quand, pour certains de ces problèmes, les données étaient divisées en groupes ou classes, on pouvait considérer qu'il s'agissait d'une deuxième variable (cette variable groupe ou classe peut être qualifiée de pseudo-variable).

La question sous-jacente était alors : est-ce que deux échantillons proviennent d'une même population ? Est-ce qu'à l'intérieur d'une même population, deux groupes d'individus distincts par une caractéristique A ont des moments (c'est-à-dire moyenne et variance) différents pour une variable X. Par exemple, les individus de moins de 90 kg sont-ils en moyenne plus petits que ceux du groupe d'individus de plus de 90 kg ou bien, les élèves au fond d'une classe ont-ils les mêmes notes que ceux des premiers rangs ?

1.1. Adéquation entre le test et la question posée

Il ne s'agit plus, comme dans la question ci-dessus, de comparer deux groupes d'individus différents, mais d'**établir un lien** (ou au contraire l'indépendance) entre deux variables quantitatives pour l'ensemble des individus (chaque individu étant caractérisé par deux variables X et Y).

• Exemple :

On veut comparer les notes d'un examen dans une classe, en milieu et en fin d'année. Peut-on, connaissant la note d'un élève à l'examen de milieu d'année, anticiper la note finale avec une précision raisonnable ?

Nous pourrons faire cette comparaison si nous démontrons l'existence d'un lien entre les deux notes à partir d'un échantillon, et prévoir alors les notes de fin d'année avec une confiance dépendante de l'intensité de cette liaison.

Notre objectif, s'il existe une relation entre les variables X et Y, sera de pouvoir donner une valeur raisonnable de Y pour un individu connaissant la valeur de X. A partir de l'exemple des notes d'une classe, il est possible d'envisager plusieurs tests et il appartient au lecteur de comprendre qu'ils ne répondent pas à la même question de départ, c'est-à-dire que les tests appropriés sont différents :

• Si nous étudions les notes d'élèves d'une classe à deux périodes de l'année, il s'agit d'une corrélation. Nous pouvons poser comme hypothèse H_0 qu'il "existe un lien entre la note obtenue par un élève en milieu et en fin d'année". Nous pourrons utiliser cette relation pour anticiper en milieu d'année, le niveau d'un élève en fin d'année et peut-être modifier l'enseignement en conséquence...

• Nous pouvons regrouper les sujets selon leurs notes en deux classes < 10 et ≥ 10, puis effectuer un test du χ^2 sur les proportions de sujets par classe pour tester l'hypothèse H_0 : "les élèves qui sont dans un des deux groupes se retrouvent dans le même en fin d'année" (variable qualitative).

• Nous pouvons regrouper les sujets selon leurs notes en deux classes < 10 et, ≥ 10 puis comparer les moyennes des notes des deux groupes en fin d'année sous l'hypothèse H_0 qu'il "n'existe pas de différence entre les résultats finaux en fonction de la note de milieu d'année" (variable quantitative).

Les deux dernières méthodes ont été abordées respectivement dans les chapitres VI et VII. On voit qu'elles induisent une "perte d'information" puisqu'au lieu d'utiliser les notes effectives de milieu d'année, on les regroupe. L'analyse de corrélation nous permettra d'utiliser toute l'information, et l'on recourra à d'autres tests quand : les conditions d'applications du test de corrélation ne sont pas remplies (on peut alors faire un χ^2), ou quand le problème posé se prête à un test plus simple (comparaison de moyennes, par exemple).

1.2. Définition des termes corrélation et régression

Pour donner une information sur la liaison entre deux variables quantitatives, on utilise les termes de *corrélation* et de *régression*. **La régression**

est une fonction $y_i=f(x_i)$ qui donne la distribution de Y conditionnellement à chaque différente valeur de X ; elle est souvent calculée dans un objectif de prévision. De façon générale, il s'agit du **lien entre deux variables quand l'une des deux n'est pas aléatoire** (on peut dire contrôlée) : la valeur x_i de la variable indépendante (ou explicative) X est fixée et il lui correspond une valeur y_i de la variable indépendante (ou expliquée) Y. La variable X peut être une valeur expérimentale ou le temps. Par exemple, on peut faire une régression du nombre de patients observés par périodes de temps ou donner l'équation de régression d'une relation dose-effet. **La corrélation nous donne l'intensité du lien entre deux variables aléatoires**, elle est symétrique, c'est-à-dire que la corrélation de Y par rapport à X est la même que celle de X par rapport à Y. Le plus souvent, la corrélation donne la mesure de l'association entre deux variables observées.

La régression permet donc de décrire une relation entre X et Y dans l'objectif de pouvoir déduire une nouvelle valeur de y_i (avec une certaine incertitude) d'après toute nouvelle valeur de la variable contrôlée x_i. La corrélation nous donne l'intensité de cette liaison : on peut donc établir le coefficient de corrélation dans tous les cas, et en particulier quand les deux variables sont aléatoires. C'est, en particulier, une statistique intéressante quand on observe "passivement" une relation entre deux variables que l'on ne contrôle pas. Précisons d'ores et déjà qu'une forte corrélation n'implique pas de relation de causalité, c'est-à-dire qu'une corrélation significative n'indique pas que la variable X produit d'elle-même un effet sur Y (une troisième variable Z peut être à l'origine des valeurs de X et Y). La mise en évidence d'une causalité repose sur des critères précis. On peut citer les critères de *Hill* (1965) ou la discussion de *Susser* (1991) à ce sujet.

L'utilisation adéquate de ces deux statistiques (corrélation et régression) sera commentée au paragraphe 4.

Nous allons établir pas à pas la méthode de calcul d'une droite de régression à partir de la formule des moindres carrés ; nous pourrons ensuite en déduire simplement la formule de la corrélation. Cette démonstration nous aidera à bien comprendre les principes d'application des tests.

2. La régression

2.1. Indépendance entre deux variables quantitatives

■ Quels types de relations peut-on observer graphiquement pour deux variables quantitatives X et Y ? Dans la pratique, on a n couples d'observations (x_i,y_i) de deux variables quantitatives X et Y, et l'on recherchera

l'existence d'une liaison entre ces deux variables. Cette relation permettra de calculer des statistiques pour X et Y qui vont résumer les observations faites sur l'échantillon. On peut visualiser cette relation sur une figure de dispersion, sous la forme d'un nuage de points comme le montrent les figures 1a et 1b ci-dessous.

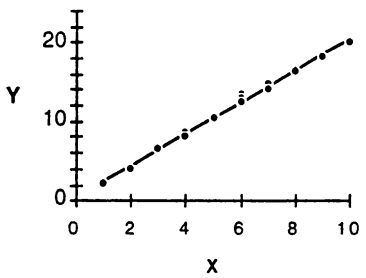


Figure 1a

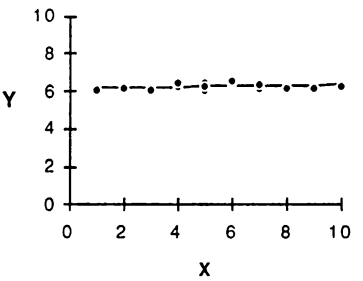


Figure 1b

Sur ce nuage de points, on peut ajuster une droite qui passe au mieux entre les points d'observation (au plus près). L'équation de cette droite est :

$$\hat{y}_i = p\,x_i + c \qquad (1)$$

où p correspond à la pente de la droite et c est une constante. Ces deux paramètres sont à estimer. On note $\hat{y}_i$ plutôt que y_i pour spécifier qu'il s'agit d'une estimation.

Nous appellerons Y la variable expliquée (ou dépendante), et X la variable explicative (ou indépendante). Les valeurs (x_i, y_i) de l'échantillon sont une réalisation de valeurs aléatoires que nous supposerons décrites par la relation linéaire. Ainsi, $\hat{y}_i = E(y_i)$ est la valeur attendue de Y sous l'hypothèse que $\hat{y}_i = p\,x_i + c$ décrit la population théorique.

• Exemple de la figure 1a :

On voit d'après cette figure que l'on trouvera une liaison positive, car à chaque valeur plus forte de X correspond une valeur différente et plus forte de Y.

• Exemple de la figure 1b :

Nous observons qu'une droite permet de modéliser la relation entre X et Y, mais que la pente de cette droite est nulle car la droite est horizontale. Quelle que soit la valeur de X, la valeur correspondante de Y est donc la même : il y a indépendance entre X et Y, ou autrement dit, la connaissance de la valeur de X ne donne aucune indication sur celle de Y (cette valeur de Y est toujours la même).

■ Dans le cas de la régression, la variable X n'est souvent pas aléatoire. La situation la plus fréquente est celle où l'on observe plusieurs valeurs de Y pour une même valeur de X fixée par exemple en recrutant les personnes d'un échantillon d'après leur âge, ou en exprimant une relation en fonction du temps (ou de périodes de temps comme le trimestre ou l'année). La relation (1) se visualise plutôt sous la forme de la figure 1c :

• *Exemple de la figure 1c :*

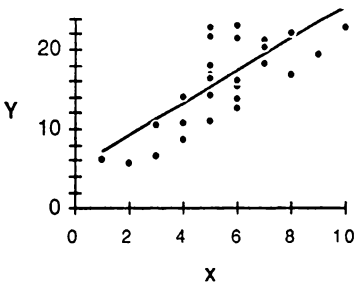


Figure 1c

Dans cet exemple, on voit qu'à chaque point d'abscisse x_i, correspondent plusieurs points y_i. Nous verrons que, dans ce type de figure, on peut concevoir quelles seront les conditions d'applications pour que la droite de régression passe au mieux entre les points d'observation et résume correctement la relation entre Y et X.

■ On peut rencontrer d'autres types de nuages de points que nous allons détailler ci-dessous :

• *Exemple de la figure 2 :*

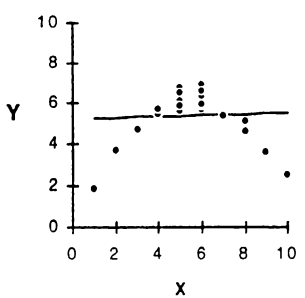


Figure 2

On voit que la relation entre X et Y n'est pas linéaire. En fait cette relation est de la forme $\hat{y} = c + p\,x + a\,x^2$, et il y a plusieurs façons de constater que l'ajustement des points par une droite n'est pas satisfaisant : le simple tracé de la figure 2 permet de constater visuellement que l'ajustement par la droite de régression n'est pas correct ; nous verrons dans le paragraphe 5 que l'analyse des résidus (écart entre le point observé et la droite) permet aussi de rejeter ce modèle linéaire.

Nous nous restreindrons dans cet ouvrage à une relation linéaire de la forme $\hat{y}_i = p\,x_i + c$; d'autres formes de régression, telles que régression non linéaire ou multiple, peuvent être formalisées. Il s'agit d'une extension des principes que nous allons aborder dans ce chapitre. Nous renvoyons pour leur étude à d'autres ouvrages tels que ceux d'Armitage (1994), Tomassonne (1983) ou Wonnacott (1995).

• *Exemple des figures 3a et 3b :*

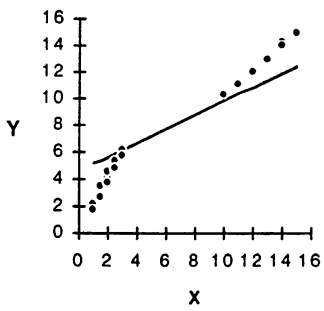


Figure 3a

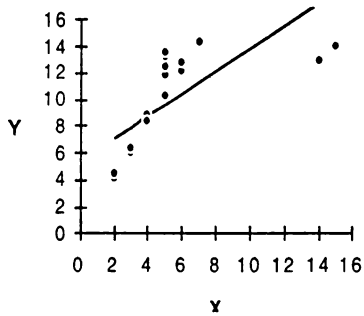


Figure 3b

Pour la figure 3a, on voit qu'il existe une relation linéaire entre les variables X et Y, mais pour deux sous-populations distinctes. La première population, pour laquelle $x_i < 4$, et la seconde population, pour laquelle $x_i > 10$, n'ont pas la même relation linéaire. On voit donc qu'il faut faire des hypothèses sur la distribution de ces variables X et Y pour éviter de se trouver dans ce cas d'application erronée.

On retrouve le même problème de distribution dans la figure 3b où il y a non pas deux sous-populations, mais seulement deux points aberrants qui faussent l'ajustement de la droite de régression. On voit que si l'on éliminait ces deux points pour les valeurs où $x_i = 14$ et $x_i = 15$, la droite de régression serait plus verticale et passerait plus correctement à travers le nuage de points : il faut donc éliminer ces valeurs aberrantes après les avoir analysées et avoir justifié le fait de les retirer de l'analyse.

Avant de trouver comment formuler ces relations par des statistiques, nous allons voir comment résumer une relation linéaire $\hat{y}_i = p\,x_i + c$ en ajustant une droite sur un ensemble de points (x_i,y_i).

2.2. Ajuster une droite au sens des moindres carrés

Pour trouver la **droite qui passe au mieux entre les points du nuage**, on utilise couramment la méthode des moindres carrés. Pour chaque observation de coordonnées (x_i,y_i), on cherche le point estimé $(x_i, \hat{y}_i)$ de la droite qui passe au plus près de l'observation. Pour ajuster la droite, il faut minimiser l'erreur de prédiction $e_i = y_i - \hat{y}_i$ (ou résidu) pour chaque observation i. C'est-à-dire qu'il s'agit de minimiser la distance entre le point observé et le point estimé. On pourrait donc chercher le minimum de la somme des écarts $|y_i - \hat{y}_i|$ (en valeur absolue pour éviter les problèmes de signe) pour obtenir la droite qui passe au plus près des n points d'observation. En fait, on va minimiser la somme des écarts au carré comme cela se fait plus couramment en statistiques, c'est le critère des Moindres Carrés (MC) :

$$MC = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Les logiciels de statistiques proposent couramment cette méthode pour trouver les paramètres p et c de la droite de régression.

2.3. Analyse de l'équation de régression

2.3.1. La pente de la droite de régression

Une fois ajustée la droite d'équation $\hat{y}_i = p\,x_i + c$ par la méthode des moindres carrés, la pente p de la droite nous montre l'existence de la relation entre X et Y (comme on l'a vu pour les graphiques 1a et 1b). On montre que la formule qui en découle pour le calcul du coefficient de régression p est :

$$p = \frac{\sum_{i=1}^{n}(y_i - m_y)(x_i - m_x)}{\sum_{i=1}^{n}(x_i - m_x)^2} = \frac{\sum_{i=1}^{n}\frac{(y_i - m_y)(x_i - m_x)}{n-1}}{\sum_{i=1}^{n}\frac{(x_i - m_x)^2}{n-1}} = \frac{\text{covariance}(x,y)}{\text{variance}(x)}$$

La quantité au numérateur est en fait la covariance de (X,Y) et celle du dénominateur est la variance de X.

■ *Covariance de X et Y :*

Dans le calcul de la pente de la droite de régression, nous avons vu qu'au dénominateur est la variance de la variable expliquée et qu'au numérateur est l'expression :

$$\text{cov}(X,Y) = \sum_{i=1}^{n} \frac{(y_i - m_y)(x_i - m_x)}{n-1}$$

Cette expression est la covariance de X et Y comme nous l'avons décrite dans le chapitre III de ce livre. Elle donne l'écart à la moyenne $y_i - m_y$ observé pour chaque écart $x_i - m_x$. Le produit des deux écarts peut donc être négatif ou positif ; il est négatif par exemple quand la variable Y observée est inférieure à la moyenne alors que X est supérieure à la moyenne. **C'est donc ce terme qui nous indique le signe de la pente p et l'existence d'une relation entre X et Y.**

Cette pente p est un accroissement, on peut l'interpréter comme la variation de Y qui correspond à la variation de X d'une unité.

On peut montrer que la droite de régression passe de toute façon par le centre de gravité du nuage de points de coordonnées (m_x, m_y), avec m_x et m_y correspondant respectivement aux moyennes des variables X et Y. En effet, quand on cherche à minimiser le critère des moindres carrés, la valeur la plus faible que l'on peut obtenir est 0 (quand la droite passe exactement par l'ensemble des points). Ceci s'écrit :

$$\sum_{i=1}^{n} (\hat{y}_i - y_i)^2 = 0$$

$$\Leftrightarrow \sum_{i=1}^{n} (\hat{y}_i - p\,x_i - c)^2 = 0$$

$$\Leftrightarrow \sum_{i=1}^{n} (\hat{y}_i - p\,x_i - c) = 0$$

$$\Leftrightarrow \sum_{i=1}^{n} \hat{y}_i - p\sum_{i=1}^{n} x_i - \sum_{i=1}^{n} c = 0$$

$$\Leftrightarrow \frac{\sum_{i=1}^{n} \hat{y}_i}{n} = p\frac{\sum_{i=1}^{n} x_i}{n} + \frac{\sum_{i=1}^{n} c}{n}$$

$$\Leftrightarrow m_y = p\,m_x + \frac{n\,c}{n}$$

$$\Leftrightarrow m_y = p\,m_x + c$$

Cette relation montre bien qu'au point d'abscisse m_x correspond celui d'ordonnée m_y par l'équation de régression, donc **la droite de régression passe par le centre de gravité du nuage de points d'observation.**

On peut en déduire la valeur de la constante dans l'équation (1) :

$$\boxed{c = m_y - p\, m_x}$$

Cette constante c est la valeur de y_i quand x_i est égal à 0.

On peut parler de prédiction, puisque l'équation $f(x) = \hat{y}_i = p\, x_i + c$ permet de prédire une valeur de y_i pour une nouvelle valeur de x_i. La prédiction de cette valeur de y_i se situera dans un intervalle de confiance comme nous l'établirons au paragraphe 2.4.

■ ***Conditions d'application de la régression :***

1) Pour éviter les mauvaises utilisations de la régression, comme dans les quelques cas que nous avons présentés au paragraphe 1, il faut que les résidus e_i soient distribués de façon aléatoire et que l'espérance de leur moyenne soit nulle $E(\hat{y} - y) = 0$ (estimation sans biais). Une figure croisant les e_i et les x_i permet de vérifier visuellement leur répartition aléatoire.

2) La distribution conditionnelle des y_i doit être de même variance pour chaque valeur de X, ce qu'on appelle une *condition d'homoscédasticité* : cette condition de conservation de la variance veut seulement que chaque valeur de X ait la même contribution dans l'ajustement de la droite de régression.

Par exemple, dans la figure ci-dessous, nous avons 30 points d'observation pour chaque valeur de X. Pour les valeurs où $x_i = 35$, on voit que la variance est plus grande et que ce sont les valeurs pour lesquelles $x_i = 15$ et $x_i = 25$ qui déterminent en grande partie l'ajustement de la droite de régression. Pour effectuer un meilleur ajustement, il est possible de rendre les données homoscédastiques par une transformation de variable adéquate telle que la transformation de Box-Cox. Il est par exemple assez courant d'utiliser le logarithme de Y quand la variance de Y augmente avec les valeurs de X.

En fait, on peut se départir de ses hypothèses dès que le nombre de points d'observations est grand ($n \geq 30$) en vérifiant de toute façon la répartition des résidus comme nous le détaillons dans le paragraphe 2.5.

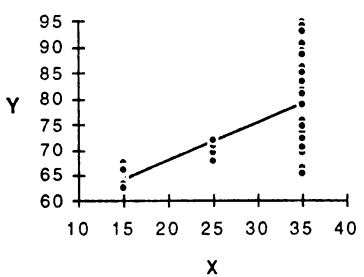


Figure 4

■ ***Remarque :***

Pour chaque nuage de points, il existe deux coefficients de régression : p_x (pente de la droite $\hat{y}_i = p\,x_1 + c$) si X est la variable contrôlée et p_y (pente de la droite $\hat{x}_i = p'\,y_1 + c'$) si Y est la variable contrôlée. Car pour la droite de régression $\hat{y}_i = p\,x_i + c$, on minimise les écarts $\hat{y}_i - y_i$, alors que pour l'autre droite $\hat{x}_i = p'\,y_i + c'$, on minimise les écarts $\hat{x}_i - x_i$. On peut le vérifier sur l'exemple suivant.

• Exemple de la figure 5 :

En étudiant la relation taille-poids chez un ensemble d'enfants, on peut sélectionner des individus d'après leur taille et observer les poids correspondants. Dans ce cas, la taille est la variable **explicative** (ou indépendante) et le poids est la variable **expliquée** (ou dépendante) ; s'il existe une droite de régression dont la pente est différente de 0, ce sera la droite de régression dès poids par rapport à la taille. Si l'on sélectionne au contraire des individus d'après leur poids pour essayer de prédire leur taille, on obtient la droite de régression de la taille par rapport au poids. On a bien deux droites de régression différentes avec, pour chaque valeur fixée de la variable explicative, une distribution aléatoire de la variable expliquée spécifique (voir figure 5) :

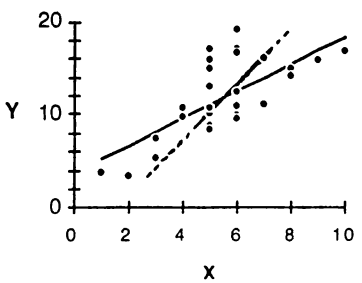


Figure 5

Dans le premier cas, on ajuste la droite de régression en minimisant la somme des écarts $\hat{y}_i - y_i$ au carré ; et dans l'autre cas on minimise la somme des écarts $\hat{x}_i - x_i$ au carré. Il est rare d'avoir à déterminer les deux droites en même temps.

2.3.2. Test sur la pente p de la droite de régression

Une fois la pente p connue, nous voulons pouvoir dire si la liaison entre les variables X et Y est significative ou non. Nous pouvons alors tester cette pente pour savoir si elle est significativement différente de 0.

Pour cela on peut se baser sur le fait que le paramètre de régression p suit une loi normale de moyenne p et de variance :

$$s_p^2 = \frac{\frac{s_y^2}{s_x^2} - p^2}{n-2}$$

avec

$$s_x^2 = \frac{\sum_i (x_i - m_x)^2}{n-1}$$

et

$$s_y^2 = \frac{\sum_i (y_i - m_y)^2}{n-1}$$

On peut alors utiliser le test de comparaison entre p et 0 avec n-2 degrés de liberté :

$$t = \frac{p-0}{\sqrt{s_p^2}}$$

soit

$$t = \frac{p}{s_p}$$

■ ***Remarque :***

On peut aussi avec ce même principe comparer deux pentes de droites de régression pour décider si elles définissent des relations différentes ou non. Il s'agit de reprendre le test t ci-dessus à l'identique d'une comparaison de deux moyennes, pour comparer deux pentes p_1 et p_2 :

$$t = \frac{p_1 - p_2}{\sqrt{s_{p1}^2 + s_{p2}^2}}$$

2.3.3. Linéarité de la droite de régression

Il est possible de **vérifier que la relation entre X et Y est linéaire quand il y a des répétitions des mesures de y pour chaque valeur x_i de X.** Nous considérerons par facilité qu'il y a j valeurs de Y_i, notées Y_{ij}, pour chaque valeur de X_i. L'équation de régression s'écrit alors différemment puisqu'il y a i fois j valeurs de Y :

$$y_{ij} = px_i + b + e_{ij}$$

■ ***Rôle du test :***

Il s'agit dans ce test de **comparer deux variances**, l'une due à la droite de régression, l'autre issue des points d'échantillonnage. Si la variance liée aux erreurs pures est importante par rapport à la variance des écarts à l'ajustement, alors on peut considérer que la relation est linéaire.

■ ***Procédure :***

Pour introduire ce test, nous allons développer un artifice de calcul sur la différence entre les points observés y_i et les points $\hat{y}_i$ estimés par la droite de régression :

$$e_{ij} = y_{ij} - \hat{y}_{ij} = (y_{ij} - m_{y_i}) + (m_{y_i} - \hat{y}_{ij})$$

$$\Rightarrow (y_{ij} - \hat{y}_{ij})^2 = (y_{ij} - m_{y_i})^2 + (m_{y_i} - \hat{y}_{ij})^2 + 2(y_{ij} - m_{y_i})(m_{y_i} - \hat{y}_{ij})$$

$$\Rightarrow \sum_{j=1}^{n}\sum_{j=1}^{n}(y_{ij} - \hat{y}_i)^2 = \sum_{j=1}^{n}\sum_{j=1}^{n}(y_{ij} - m_{y_i})^2 + \sum_{j=1}^{n}\sum_{j=1}^{n}(m_{y_i} - \hat{y}_{ij})^2 + 0$$

On peut résumer cette relation par la formule :

$$\boxed{\sum \text{des résidus} = \sum \text{erreurs pures} + \sum \text{écarts à l'ajustement}}$$

où $\sum$ est la somme des écarts au carré.

Si l'on a k valeurs de X avec n_j répétitions pour chacune d'elles, alors la variance résiduelle a n-2 degrés de liberté, et celle de l'erreur pure en n-k. La variance des écarts à l'ajustement a donc k-2 degrés de liberté. On peut donc calculer le rapport des variances pour tester la linéarité de la relation :

$$\boxed{F_{n-k}^{k-2} = \frac{\dfrac{\sum_{j=1}^{n}\sum_{j=1}^{n}(m_{y_i} - \hat{y}_{ij})^2}{k-2}}{\dfrac{\sum_{j=1}^{n}\sum_{j=1}^{n}(y_{ij} - m_{y_i})^2}{n-k}}}$$

Ce test n'est donc pas toujours réalisable, mais il peut être mis en œuvre quand l'expérimentateur contrôle la variable X et peut échantillonner plusieurs valeurs y_{ij} pour chaque x_i.

2.4. Intervalles de confiance

2.4.1. Intervalle de confiance de la droite de régression

On perçoit intuitivement que, lorsque nous ajustons une droite de régression, la confiance que nous accordons à cet ajustement dépend de la

dispersion des points autour de la droite de régression. Plus cette dispersion est grande, plus l'intervalle de confiance autour de la droite de régression sera large :

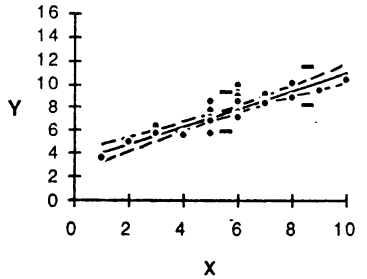


Figure 6a

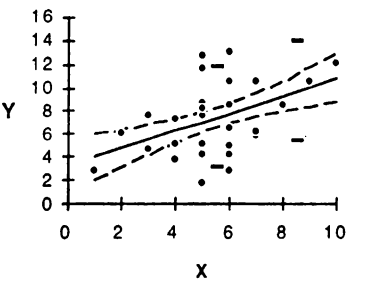


Figure 6b

L'intervalle de confiance de la droite de régression est tracé en pointillé sur les figures 6a et 6b ci-dessus. Cet intervalle de confiance autour de la droite de régression est celui dans lequel doit se trouver la droite théorique "que nous ne connaissons pas" mais que nous estimons. Cela peut aussi être exprimé en disant qu'avec un autre échantillon, on aurait trouvé une droite de régression différente, mais dont le tracé devrait être compris entre ces deux intervalles au risque α. Cet intervalle s'écrit :

$$\hat{y}_i - \varepsilon s_e \sqrt{\frac{1}{n} + \frac{(x_i - m_x)^2}{\sum_{j=1}^{n}(x_j - m_x)^2}} \leq \hat{y}_i \leq \hat{y}_i + \varepsilon s_e \sqrt{\frac{1}{n} + \frac{(x_i - m_x)^2}{\sum_{j=1}^{n}(x_j - m_x)^2}}$$

où s_e est la variance des résidus :

$$s_e = \sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n}$$

et ε est lu dans la table de la loi normale.

On constate que l'intervalle de confiance augmente quand on s'écarte du centre de gravité de coordonnées (m_x, m_y). C'est pourquoi il faut éviter de tracer la droite de régression au-delà des points d'observation car cela peut conduire à une utilisation abusive de la relation obtenue.

2.4.2. Prévision pour un nouveau point d'abcisse x_j

On veut connaître l'intervalle de confiance, non plus autour de la droite des points $\hat{y}_i$, mais autour d'un point seul d'observation $\hat{y}_i$. Dans la pratique, on connaît la relation entre $\hat{y}_i$ et x_j, et l'on veut prédire les

valeurs probables de $\hat{y}_i$ quand on a fixé une nouvelle valeur non observée de x_i. On obtient l'intervalle de confiance d'un point au risque $\alpha = 5\ \%$, qui est figuré par les traits épais sur les figures 6a et 6b précédentes pour $x_i = 5{,}5$ et $s_2 = 8{,}5$; il est plus large que l'intervalle de confiance autour de la droite de régression.

2.5. Analyse des résidus

2.5.1. Vérification des conditions d'application

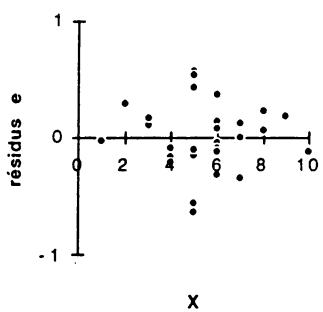
Un résidu est la valeur de la différence entre un point observé et la droite ajustée. Nous l'écrivons $e_i = \hat{y}_i - y_i$, puisqu'il s'agit de l'écart entre le point d'observation et la droite. L'équation de régression s'écrit alors :

$$\hat{y}_i = p\,x_i + c + e_i$$

dans laquelle e_i est l'erreur d'estimation sur la $i^{\text{ème}}$ observation.

L'analyse des résidus est une phase importante pour décider si le modèle linéaire est convenable. On peut en juger sur la figure 3a, les résidus vont être négatifs pour $x_i < 4$, puis positifs après : les résidus ne sont pas répartis de façon aléatoire autour de la droite de régression qui ne passe donc pas correctement entre les points d'observation.

Il est d'usage de tracer un graphique croisé entre les x_i et les résidus e_i pour vérifier que leur répartition est aléatoire. On peut comparer le résultat obtenu à partir des régressions des figures 2 et 3 qui n'étaient pas faites en respectant les conditions d'application, et celui de la figure 5 qui semble les respecter.



Résidus de la figure 5

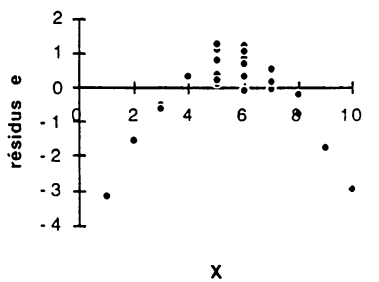
L'observation de ces figures permet de visualiser rapidement comment les résidus sont répartis autour de 0, autrement dit comment la droite de

régression passe entre les points d'observation. Dans le graphique ci-dessus, des résidus de la figure 5, on voit que ceux-ci sont répartis aléatoirement et de façon symétrique autour de la droite 0 (la distribution des résidus est même normale) ; on ne pourra pas améliorer la relation entre les valeurs de X et Y, sauf peut-être en variance puisque l'on voit que la variance de Y est plus grande (i.e. les écarts e sont plus importants) pour des valeurs moyennes croissantes de X.

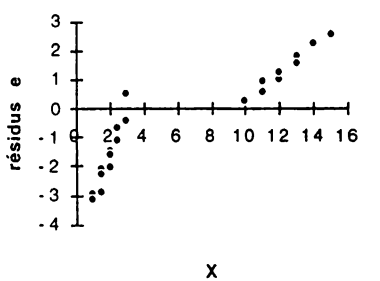
Donc, nous ne sommes pas exactement dans les conditions d'hétéroscédasticité, mais on peut accepter les résultats de l'étude de régression comme convenables. Il aurait fallu revoir l'ajustement de la droite si la variance de Y augmentait avec les valeurs de X (et dans ce cas les résidus seraient de plus en plus écartés de 0 quand X tend vers 10). Si l'on n'est pas dans les conditions d'homoscédasticité, on peut alors ajuster la droite de régression sur une transformation de Y (comme le logarithme, qui est une restriction de la transformation de Box-Cox) ou bien effectuer une régression pondérée qui donne plus de poids dans l'ajustement aux valeurs de faible variance (ici elles correspondent aux valeurs les plus faibles de X). Dans ce dernier cas, l'ajustement est fait en minimisant un moindre carré pondéré (MCP) par la variance comme dans la formule suivante :

$$MCP = \sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{s_i^2}$$

Les deux exemples donnés dans les figures 2 et 3 ne respectaient pas à l'évidence les conditions d'application de l'ajustement d'une droite de régression linéaire. On peut l'observer tout aussi immédiatement sur les figures des résidus ci-dessous :



Résidus de la figure 2



Résidus de la figure 3

Dans la figure 2, les résidus sont presque répartis symétriquement par rapport à 0 mais peuvent encore être décrits par une relation non linéaire

qu'il faudra réintroduire dans l'équation de régression. Cela peut se juger aussi sur la figure 3 où l'on trouve une série de résidus négatifs, puis une autre série de résidus positifs. Le mauvais choix du modèle est ici assez facile à constater, mais dans des situations moins extrêmes, il est commode de pouvoir disposer d'un test sur la répartition aléatoire des résidus. Le test que nous proposons ici est le test des séquences (qui dans ces deux derniers cas montrerait que les résidus ne sont pas aléatoires).

2.5.2. Test des séquences sur les résidus

■ *Rôle du test :*

Le rôle de ce test est de **détecter un ajustement qui ne passe pas correctement au milieu des points d'observation**, mais parfois en dessous ou au-dessus d'un groupe de points. Le principe consiste à compter les séries de résidus négatifs ou positifs et d'établir un score qui nous dise s'il y a alternance "aléatoire" entre les résidus positifs et négatifs (donc si la droite de régression passe presque alternativement en dessous et au-dessus des points d'observation), ou bien si au contraire elle passe trop longtemps en dessous et au-dessus avec alors des séries de résidus négatifs ou positifs.

■ *Procédure :*

Pour réaliser ce test, on va compter le nombre de séquences S (une séquence étant une suite e_i, e_{i+1},...,e_{i+s} de résidus de même signe) entre x_0 et x_n. Pour les résidus de la figure 3 présentés dans le paragraphe précédent, ce calcul est assez simple puisque les 11 premiers résidus, quand x_i va de 0 à 3, sont négatifs (première séquence) et les 19 autres sont positifs (deuxième séquence) : on trouve S = 2. On voit que si la droite de régression passe bien entre les points d'observation, les résidus seront alternativement positifs, puis négatifs ; dans ce cas les séquences seront de taille 1 ou 2 à chaque fois. On montre que le nombre de séquences S attendues sous l'hypothèse H_0 (selon laquelle les résidus sont répartis de façon aléatoire autour de 0) est :

$$E(S) = \frac{n}{2} + 1$$

avec une variance :

$$\text{var}(S) = \frac{n-1}{4}$$

On peut donc tester si le nombre de séquences S observées est significativement différent de celui attendu sous l'hypothèse H_0 :

$$t = \frac{S - E(S)}{\sqrt{var(S)}}$$

à comparer à la valeur de la table de la loi normale.

Il existe aussi un test de Durbin-Watson qui permet de vérifier si les résidus e_i et e_{i-1} sont indépendants (ce que l'on appelle l'autocorrélation). Si l'on trouve d'après ce test que les résidus sont autocorrélés entre eux, c'est qu'il existe une relation entre eux qui aurait pu être introduite dans l'équation de régression.

2.5.3. Distribution des résidus

Nous avons dit que **la distribution de la variable Y devait être normale** : cela implique que **les résidus aient une distribution normale autour de 0.** On peut tester si l'on se trouve bien dans les conditions d'application en vérifiant que les résidus ont une distribution normale autour de 0. Si ces résidus sont distribués normalement autour de 0, alors on peut considérer que la variable Y est distribuée normalement autour de $p\,m_x+c$. En effet, les résidus e_i et les valeurs observées Y_i sont liés par une relation déterministe (non aléatoire) $y_i = p\,x_i + c + e_i$. Ceci permet donc de vérifier *a posteriori* que l'on se trouve dans les conditions d'application de la régression linéaire.

3. La corrélation

La corrélation reprend les mêmes principes que ceux abordés pour la régression : **il s'agit d'ajuster une droite sur un ensemble de points d'observation (x_i,y_i) et de l'utiliser pour donner une statistique sur l'existence voire l'intensité de la liaison entre les deux variables aléatoires.** Il peut s'agir par exemple de l'étude de la relation poids-taille pour un échantillon d'individus tirés au sort.

Dans le cas de la corrélation, on ne contrôle aucune des deux variables, ce qui fait qu'on peut aussi bien rechercher le critère des moindres carrés pour la relation $\hat{x} = \frac{1}{p}\,y - \frac{c}{p}$, en renversant l'équation (1). Nous avons vu qu'il existait deux droites de régression selon que l'on prenne X ou Y comme variable dépendante. Cette fois, chaque variable ayant un rôle symétrique, nous voulons ajuster une seule droite $\hat{y}_i = p\,x_i + c + e_i$ sur l'ensemble des points d'observation identiques à celle que nous aurions trouvée avec l'autre équation possible $\hat{x} = \frac{1}{p}\,y - \frac{c}{p}$: nous avons déjà abordé un problème similaire de la standardisation d'une variable dans le cas de la loi normale centrée-réduite.

3.1. Le coefficient de corrélation de Bravais-Pearson

3.1.1. Rôle du test de Bravais-Pearson

Montrer l'existence d'une liaison linéaire entre deux variables quantitatives et déterminer l'intensité de cette liaison.

Hypothèses :

H_0 : on suppose l'indépendance linéaire entre les variables X et Y ($\rho = 0$)

H_1 : existence d'une relation linéaire entre les variables X et Y ($\rho \neq 0$)

3.1.2. Procédure

Nous connaissons une façon efficace de standardiser une variable : nous avons vu que l'**on peut standardiser une loi normale en soustrayant par la moyenne et en divisant par la variance.** Nous pouvons appliquer le même principe en divisant Y par sa variance et X par sa variance ; ainsi l'unité utilisée ne joue plus sur la pente de la droite de régression puisqu'elle est standardisée. Par exemple, dans la relation taille-poids, si pour une même unité de poids (en kg), la taille est exprimée en mètres ou en centimètres, la pente de la droite ne sera pas la même car la dispersion des valeurs en centimètres sera plus grande par rapport aux valeurs de poids en kg. Il faut donc que chaque variable X et Y soit exprimée dans une échelle standardisée et il suffira de diviser chaque valeur observée par la variance. C'est ce que l'on écrit dans la relation suivante pour déduire l'équation de la corrélation r :

$$\frac{y}{s_y} = r\,\frac{x}{s_x} + c \Leftrightarrow y = r\,\frac{s_y}{s_x}\,x + c$$

d'où

$$p = r\,\frac{s_y}{s_x} \Leftrightarrow r = p\,\frac{s_x}{s_y}$$

r est le coefficient de corrélation de Bravais-Pearson :

$$r = p\frac{s_x}{s_y} = \frac{s_{xy}^2}{s_x^2}\frac{s_x}{s_y}$$

$$\text{soit, } r = \frac{s_{xy}^2}{s_x s_y} r = \frac{\text{covariance}(x,y)}{\sqrt{\text{variance}(x) \times \text{variance}(y)}}$$

On peut, pour le calculer, réécrire cette expression du coefficient de corrélation r comme :

$$r = \frac{\sum_{i=1}^{n} \frac{(y_i - m_y)(x_i - m_x)}{n}}{\sqrt{\sum_{i=1}^{n} \frac{(y_i - m_y)^2}{n} \sum_{i=1}^{n} \frac{(x_i - m_x)^2}{n}}}$$

S'agissant d'un rapport de variance, **cette valeur de r est toujours comprise entre -1 et +1. Elle exprime la pente de la droite des moindres carrés en coordonnées réduites**. Elle est donc symétrique, au contraire du coefficient de régression p, c'est-à-dire que r est identique, que l'on obtienne l'ajustement par les moindres carrés d'après la droite $\hat{y} = p\,x + c$ ou la droite $\hat{x} = p'\,y + c'$.

3.1.3. Lecture de la table du coefficient de corrélation

La table du coefficient de corrélation, présentée en annexe, nous donne la **probabilité pour que le coefficient r calculé égale ou dépasse** (en valeur absolue) **une valeur donnée avec un degré de liberté égal à n-2.**

Par exemple, la valeur $\alpha = 0{,}01$ lue dans la table de la loi normale à n-2 = 20-2 = 18 degré de liberté est r = 0,5614. Autrement dit, la probabilité pour que le coefficient de corrélation calculé r soit supérieur à 0,56 (ou inférieur à -0,56) est inférieure à 1 %.

Si 20 points d'observation passent tous par la droite y = x, alors le coefficient de corrélation sera r = 1 comme figuré sur la figure 7 ci-dessous ; si ces 20 points ne sont pas aussi bien alignés autour de la droite y = x (ce qui est plus réaliste) alors le coefficient de corrélation sera inférieur à 1. Si ce coefficient est r = 0,5 pour n = 20, on a une probabilité inférieure à 5 % ($r_{28ddl\ (5\ \%)} = 0{,}4438$) d'observer ce r supérieur à 0,44 et l'on pourra conclure à une bonne intensité de la liaison entre X et Y. Si le coefficient est r = 0,4, alors on a plus de 5 % de chance d'obtenir ce résultat et l'on peut conclure que la droite décrit mal l'intensité de la relation entre X et Y.

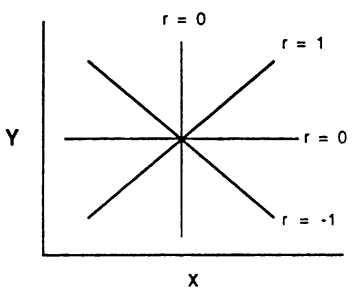


Figure 7

■ *Commentaires sur la figure 7 :*

La figure 7 schématise plusieurs cas possibles.

Si la relation est une droite horizontale, alors la corrélation est nulle puisqu'à chaque valeur de x_i correspond une même valeur de y_i. De même, si la relation est une droite verticale alors il y a aussi indépendance entre X et Y, et la valeur de x_i est la même quel que soit y_i. Dans les deux autres cas, la relation entre X et Y existe ; elle peut être positive comme nous l'avons représenté dans le graphique 1a ou bien négative dans l'autre cas. Autrement dit, quand elle est positive, la relation linéaire indique que plus x_i augmente, plus y_i augmente aussi ; quand elle est négative, la relation indique qu'à chaque valeur croissante de x_i correspond une valeur plus faible de y_i.

3.1.4. Rôle et variations du coefficient de corrélation

Le coefficient de corrélation mesure l'intensité de la relation linéaire entre X et Y, c'est-à-dire que **plus la valeur du coefficient r est proche de 1, plus les points d'observation (x_i,y_i) sont proches de la droite de régression.**

Par exemple, le coefficient de corrélation peut être identique pour les figures 1a et 1b. Si le coefficient de corrélation r est proche de 1, alors la droite décrit très bien le nuage des points d'observation ; au contraire quand r tend vers 0 (se rapproche de 0), c'est que la droite résume assez mal la distribution du nuage de points (autrement dit les écarts entre la droite et les points observés sont grands). **Pour vérifier l'indépendance linéaire entre les deux variables, ce qu'il faut tester c'est la pente de la droite** afin de déterminer si la relation existe (comme dans la figure 1a), ou non (comme dans la figure 1b).

On peut argumenter le rôle du coefficient de corrélation à mesurer l'intensité de la relation linéaire entre X et Y. Si les points d'observation (x_i,y_i) sont proches de la droite de régression (donc relation forte comme dans la figure 1a), alors la variance liée $\sigma^2_{y/x}$ de Y par rapport à X est faible. Autrement dit, pour chaque valeur de x fixée, les valeurs de y correspondantes sont distribuées normalement autour de $\hat{y}$ avec une variance $\sigma^2_{y/x}$ faible. On peut alors juger du rôle de ρ (la valeur théorique du coefficient de corrélation), sachant que la variance marginale σ^2 de y est proportionnelle à la variance liée $\sigma^2_{y/x}$ d'un facteur $(1-\rho^2)$:

$$\sigma^2_{y/x} = \sigma^2_y(1-\rho^2)$$

On voit alors, qu'à variance marginale constante, si le coefficient de corrélation ρ augmente, alors la variance liée $\sigma^2_{y/x}$ diminue, donc les points

d'observation (x_i,y_i) sont proches de la droite de régression, ou autrement dit l'intensité de la relation entre X et Y est forte.

Posons par exemple $\rho = 0$, alors l'intensité de la relation est faible puisque la variance liée $\sigma^2_{y/x}$ est égale à la variance marginale σ^2. Si $\rho = 1$, alors l'intensité de la relation est forte puisque la variance liée est nulle (tous les points d'observation (x_i,y_i) sont sur la droite de régression).

Renversons aussi cette relation pour montrer que le carré du coefficient de corrélation (coefficient de détermination) est le rapport de la variance liée sur la variance de Y (à une constante près) :

$$\sigma^2_{y/x} = \sigma^2_y(1-\rho^2)$$

soit,
$$\frac{\sigma^2_{y/x}}{\sigma^2_y} = 1-\rho^2$$

soit,
$$\rho^2 = \frac{\sigma^2_{y/x}}{\sigma^2_y} - 1$$

Ainsi, on comprend mieux le coefficient de détermination comme une fraction de la variance expliquée par la droite de régression.

3.2. Test et intervalle de confiance

On peut prendre comme hypothèse H_0 que la pente ρ est nulle (pas de relation linéaire entre les deux variables) $\rho = 0$, et tester cette hypothèse. On compare alors le coefficient r obtenu à celui de la table du coefficient de corrélation au risque α avec un nombre de degré de liberté n-2 (n étant le nombre d'observations). Pour n = 102 observations, on peut parier avec un risque d'erreur $\alpha = 5$ % que la valeur de |r| sera supérieure à 0,1946, lu dans la table, s'il existe une relation linéaire significative entre Y et X.

Il n'est pas toujours possible de trouver la valeur du risque dans les tables du coefficient de corrélation ; dans ce cas on peut utiliser la table de t en recalculant la valeur correspondante :

$$t = \frac{r}{\sqrt{1-r^2}}\sqrt{n-2}$$ avec le même nombre de degrés de liberté n-2.

Ce test se déduit de celui de la pente de la régression comme nous le montrerons dans l'exercice 2 de ce chapitre.

■ ***Conditions d'application :***

Il faut qu'une des deux distributions au moins soit normale.

On a vu sur les figures 3 que si la distribution s'écarte d'une forme

normale, cela peut amener à des nuages de points avec des formes particulières : pour la distribution 3b, on voit qu'un seul point a une très forte contribution à lui seul sur la pente de la droite et que cela ne reflète pas correctement la forme de la relation ; pour la distribution 3, on peut penser qu'il y a deux sous-populations distinctes, il faut le tester et étudier les corrélations séparément si le nombre de sujets est suffisant.

■ *Remarque :*

Notons bien que la réalisation d'un test, conduisant à la conclusion d'une différence non significative de r à la valeur 0, ne veut pas dire que X et Y sont indépendants. Il peut très bien exister une relation non linéaire entre X et Y, comme par exemple une relation en forme de courbe en U ou en cloche.

3.3. La corrélation sur les rangs

3.3.1. Rôle du test de corrélation sur les rangs

Dans certains cas, il n'est pas possible d'appliquer le test du coefficient de corrélation de Bravais-Pearson. Il existe des **tests de corrélation sur les rangs** qui représentent une alternative quand on dispose de **mesures qualitatives ordonnées** ou qu'on ne peut pas supposer que la distribution est normale.

Hypothèses :

H_0 : l'ordre des x_i est indépendant de celui des y_i ($\rho' = 0$)
H_1 : l'ordre des x_i est lié à celui des y_i ($\rho' \neq 0$)

3.3.2. Procédure

Le test de Spearman que nous présentons ci-dessous est un test non paramétrique. Il consiste à remplacer les valeurs de X et Y par leurs rangs dans le calcul proposé par Bravais et Pearson.

On remplace donc chaque valeur de x_i par son rang parmi x_i ; en cas d'ex aequo, on tire au sort l'ordre des ex aequo.

On peut montrer que, dans ce cas, la formule de Bravais-Pearson se simplifie par le calcul du coefficient ρ' de corrélation de Spearman :

$$\rho' = 1 - 6\frac{\sum_{i=1}^{n} d_i^2}{n(n^2-1)}$$

avec $d_i = y_i - x_i$

On peut comparer le coefficient obtenu à celui de la table du coefficient de Spearman en annexe, au degré de liberté n-2. Si $n \geq 10$, alors on convient qu'il est possible d'utiliser la table du coefficient de corrélation de Bravais-Pearson.

Le coefficient ρ' de Spearman ne donne pas l'intensité de la liaison quand il tend vers 1, **il permet seulement de déterminer s'il existe ou non une liaison linéaire entre les deux variables X et Y.** Ce test non paramétrique peut être utilisé pour confirmer les résultats obtenus avec le coefficient de Bravais-Pearson quand on n'est pas sûr d'être dans les conditions d'application.

Le test est le même que celui de Bravais-Pearson :

$$t = \frac{\rho}{\sqrt{1-\rho^2}}\sqrt{n-2}$$

avec n-2 degré de liberté, en suivant la procédure du test t du chapitre VII.

Notons qu'il existe un autre test de corrélation des rangs développé par Kendall (1990).

4. Utilisation pertinente de la régression et/ou corrélation

Nous avons décrit, dans ce chapitre, l'intérêt des calculs de régression et corrélation en précisant qu'il était possible de fournir ces deux résultats dans l'analyse d'une liaison entre deux variables quantitatives X et Y.

Il faut cependant faire certaines restrictions sur les différentes utilisations de ces notions. La droite de régression peut être établie dans tous les cas d'analyse et ce d'autant qu'elle permet de vérifier la distribution des résidus. Mais si aucune des deux variables ne peut être contrôlée, cela n'apporte pas grand-chose au résultat puisque cette droite est surtout utile pour prédire d'autres valeurs. On pourrait cependant trouver une justification quand il existe des valeurs manquantes pour une des deux variables et qu'on cherche à les extrapoler.

La corrélation peut être donnée en même temps que l'équation de la droite de régression, mais il faut faire attention car le coefficient de corrélation peut varier en fonction de l'étendue de la variable contrôlée X. Si l'on fixe certaines valeurs de X pour calculer la droite de régression sur un intervalle plus faible que celui qu'aurait donné un tirage aléatoire des valeurs de X, alors le coefficient de corrélation sera plus faible (même si la droite de régression est identique à celle obtenue avec un tirage aléatoire). A l'inverse, si l'on prend des valeurs de X sur un grand intervalle, alors les valeurs extrêmes vont tendre à augmenter la valeur du coefficient de corrélation et l'intensité de la liaison sera alors plus forte.

On peut vérifier cela en tirant au sort des individus afin d'établir la relation entre l'année de naissance et la taille à l'âge adulte. Imaginons que les valeurs de X pour l'échantillon soient comprises entre 1930 et 1970, si l'on refait la même analyse en sélectionnant les individus d'après leur date de naissance et que celles-ci soient comprises entre 1950 et 1960, alors le coefficient de corrélation sera plus faible même si la droite de régression est identique. **En fait, la corrélation a des restrictions d'utilisation, car la variable X doit être une variable aléatoire ; la régression a donc un champ d'application plus large.**

On voit donc que **la corrélation s'applique mieux à l'étude de variables aléatoires non contrôlées** (disons observées) dans laquelle un tirage au sort et la taille de l'échantillon permettent de conserver une certaine reproductibilité des résultats. Il nous paraît donc abusif d'opposer ces deux notions comme cela peut se trouver dans certains ouvrages ; il s'agit de deux statistiques complémentaires qui n'ont pas toujours lieu d'être présentées simultanément et dont l'utilisation dépend de la question posée.

Dans la pratique, on peut toujours ajuster une droite au sens des Moindres Carrés entre deux variables quantitatives pour définir la relation qui les lie. D'autres livres vous décriront les régressions non linéaires (quand la forme de cette relation n'est pas linéaire), les analyses de régression multiples (quand plusieurs variables, dites explicatives, sont liées à la variable Y). On précise généralement dans le même temps l'intensité de ces liaisons en calculant le coefficient de corrélation.

Par ailleurs, quand on ne recherche pas de relation de causalité comme celle qu'implique une prévision par la méthode de régression, on peut se contenter d'établir les corrélations entre plusieurs variables. En particulier, quand on a un grand nombre de variables quantitatives pour un échantillon, on effectue des **analyses factorielles** qui permettent de regrouper les variables les plus liées (c'est-à-dire avec les corrélations les plus fortes) et de rechercher d'éventuelles relations causales par la suite. Par exemple, si l'on veut comprendre ce qui détermine la taille à l'âge adulte, on peut recueillir des informations sur celle des parents, la taille à la naissance, le poids à différents âges, la durée de la grossesse de la mère, l'alimentation, l'équilibre familial, la situation géographique, les variables environnementales, etc. On constatera alors que la corrélation existe entre la taille adulte et celle des parents, ce qui permettra ensuite de dégager des hypothèses sur les facteurs déterminants de la taille adulte et d'essayer ensuite de démontrer la relation de causalité. La bonne compréhension du calcul de la corrélation vous permettra ensuite d'aborder plus aisément les méthodes de traitement d'un grand nombre d'informations comme **l'analyse en composante principale**, les méthodes de classification, etc.

5. Exercices

5.1. Exercices corrigés (réponses pages 312 à 315)

• Exercice 1 :

Dans un échantillon de taille n = 1 000 pères ayant chacun un seul fils, on s'intéresse à la relation entre la taille du père et celle du fils. On observe que la taille du père (variable x) comme celle du fils (variable y) ont une distribution normale identique de moyenne $m_x = m_y = 1{,}75$ m et d'écart-type $s_x = s_y = 7$ cm.

1. Le coefficient de corrélation calculé est r = 0,65. Quelle est la taille moyenne des fils dont le père mesure 190 cm ?

2. Peut-on conclure à une évolution de la taille moyenne des hommes au cours du temps (sur une génération) ?

• Exercice 2 :

On veut tester si l'ajustement d'une droite de régression est correct en vérifiant que les résidus sont répartis aléatoirement autour de 0. Dans la figure 3, le nombre de séquences observées est 2 et le nombre d'observations n = 30. Concluez sur la qualité de l'ajustement obtenu.

• Exercice 3 :

Montrez qu'une droite au sens des moindres carrés passe toujours par le barycentre du nuage de points d'observation, c'est-à-dire le point de coordonnées (m_x, m_y). Autrement dit, montrer à partir de l'équation de la droite de régression, que le point d'abscisse m_x a pour ordonnée la valeur m_y.

• Exercice 4 :

La valeur d'un coefficient de corrélation pour un échantillon de taille n = 18 est r = 0,39.

1. Etablissez si ce coefficient est significativement supérieur à 0 au seuil de signification de :

   a. 5 %,

   b. 1 %.

2. Quel est le nombre d'invidus nécessaire pour montrer une différence significative à 1 % en supposant que le coefficient de corrélation est supérieur à 0 (on ne tiendra pas compte du problème de puissance) ?

5.2. Exercices non corrigés

• *Exercice 5 :*

On s'intéresse au nombre de suicides en fonction de la taille des villes en supposant qu'il y a une relation linéaire entre les deux. On recueille les statistiques de décès pour 30 villes tirées au sort parmi celles de plus de 10 000 habitants et l'on se propose d'engager des études dans ces villes. On obtient le résultat suivant pour l'année 1990 :

population	102000	102400	105500	108000	108000	109900	115600	119200	119800	123900	126900	133400	133600	136200	136400
decès en 1990	10019	9393	8154	10898	8509	10606	9028	13919	10069	14948	10240	12217	15717	14479	10473
(suite)															
population	140200	146200	148500	151600	153100	154000	170200	178300	185200	197200	201600	203500	210900	213300	252000
decès en 1990	14591	11004	15429	13069	15052	13240	15259	16345	20140	18026	16363	18750	20260	23158	24451

1. Calculez l'équation de la droite de régression du nombre de suicides en fonction de la taille des villes.
2. Testez si la pente de la régression est différente de 0 et concluez.
3. Quelle est la variance expliquée par la régression ? testez si la relation est linéaire ?
4. Tracez la courbe, vous constaterez que la relation n'est pas linéaire aux extrêmes. Conservez-vous les mêmes conclusions ? Comment aborderiez-vous le problème ?

• *Exercice 6 :*

En reprenant l'exercice 1, on remarque qu'il n'est pas nécessaire d'utiliser la formule pour calculer le coefficient de corrélation, mais simplement la relation $r = p\frac{s_y^2}{s_x^2}$.

Si l'on teste le coefficient de corrélation, quel résultat obtient-on ?
Pour retrouver ce résultat, démontrez que le test t est le même quand il s'agit de tester la pente de la droite de régression, ou de tester le coefficient de corrélation.
Partez de la formule : $t = \frac{p}{\sqrt{s_p^2}}$

5.3. Solutions

• Correction de l'exercice 1 :

1. Pour calculer la taille des fils, on recherche l'équation de la pente p de la droite de régression en utilisant la relation

$r = p\frac{s_y^2}{s_x^2}$ d'où $p = r\frac{s_x^2}{s_y^2}$: on trouve $p = 0{,}65\frac{49}{49} = 0{,}65$.

Attention : la taille était donnée en mètres et la variance en cm ; nous présentons les calculs en uniformisant les mesures en cm.

On peut déduire la valeur de b, en utilisant le fait que la droite de régression passe par le point de coordonnées (m_x, m_y) d'où :

$$175 = 0{,}65 \times 175 + b$$
$$\Leftrightarrow b = 175 \times (1 - 0{,}65)$$
$$\Leftrightarrow b = 61{,}25$$

L'équation s'écrit donc $\hat{y}_i = 0{,}65x_i + 61{,}25$. Si la taille du père est $x_i = 190$ cm, on trouve alors la taille du fils : $\hat{y}_i = 0{,}65 \times 190 + 61{,}25 = 184{,}75$. Le fils est donc plus petit que le père.

2. Il ne faut pas conclure à une évolution de la taille des hommes au cours du temps puisqu'il a été précisé que la taille moyenne reste constante entre les deux générations ($m_x = m_y = 1{,}75$ m). Ce résultat correspond à ce que Francis Galton (1822-1911) avait appelé la régression vers la moyenne ; il avait observé que les pères dont la taille est supérieure à la moyenne avaient des fils en moyenne plus petits qu'eux avec une taille se rapprochant de la moyenne de leur génération, ce qu'il a appelé "régression" vers la moyenne.

• Correction de l'exercice 2 :

Sachant qu'il y a 30 observations, on peut calculer le nombre attendu $E(S) = \frac{n}{2} + 1 = 16$ avec une variance $var(S) = \frac{n-1}{4} = 7{,}25$.

D'autre part, le nombre de séquences observées est 2, on obtient donc le test :

$$t = \frac{2 - 16}{\sqrt{7{,}25}} \approx -\frac{14}{2{,}69} \approx -5{,}2$$

Pour $P(z \leq -5{,}2)$, on trouve 0,0000000579 dans la table de la loi normale, c'est-à-dire $5{,}79\ 10^{-8}$. On peut donc rejeter l'hypothèse H_0 au risque $\alpha = 5\ \%$ avec un degré de signification inférieur à 10^{-7} (inférieur à 0,0000001), et considérer que les résidus ne sont pas répartis aléatoirement. Ceci nous conduit alors à revoir les conditions d'application pour corriger l'ajustement de la droite de régression.

• Correction de l'exercice 3 :

Si l'on somme toutes les n valeurs de Y pour l'échantillon, on obtient :

$$\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} px_i + b$$

$$\Leftrightarrow \sum_{i=1}^{n} y_i = p\sum_{i=1}^{n} x_i + \sum_{i=1}^{n} b$$

$$\Leftrightarrow \sum_{i=1}^{n} y_i = p\sum_{i=1}^{n} x_i + nb$$

En divisant par n, on retrouve les formules de la moyenne de X et de Y :

$$\Leftrightarrow \frac{\sum_{i=1}^{n} y_i}{n} = \frac{p\sum_{i=1}^{n} x_i}{n} + \frac{n}{n}b$$

$$\Leftrightarrow \frac{\sum_{i=1}^{n} y_i}{n} = p\frac{\sum_{i=1}^{n} x_i}{n} + b$$

$$\Leftrightarrow m_y = pm_x + b$$

On a montré que le point d'abscisse m_x a pour ordonnée la valeur m_y, donc que la droite au sens des moindres carrés passe par le barycentre du nuage de points.

• Correction de l'exercice 4 :

1. Les hypothèses sont les suivantes :

$$\begin{cases} H_0 : \rho = 0 \\ H_1 : \rho > 0 \end{cases}$$

a. Le test de Student à envisager est unilatéral, puisque l'hypothèse alternative H_1 écarte la possibilité que ρ soit inférieur à 0. On a pris comme risque $\alpha = 5\ \%$ que ρ soit supérieur à 0, donc il faudra comparer le t calculé à celui de t au risque $\alpha = 10\ \%$ car la table de t donne la probabilité que t soit égal ou dépasse une valeur donnée en valeur absolue (t supérieur ou égal à la valeur seuil, et -t inférieur ou égal à la valeur seuil). On trouve dans la table $t = 1{,}746$.

Calculons le t correspondant à l'échantillon selon la formule :

$$t = \frac{r}{\sqrt{1 - r^2}}\sqrt{n-2}$$

$$t = \frac{0,39}{\sqrt{1-0,39^2}}\sqrt{18-2} = \frac{0,39}{\sqrt{0,8479}}4 = 1,694$$

On ne peut donc pas rejeter l'hypothèse H_0 puisque $t < 1,746$.

b. Si l'on ne peut pas conclure à une différence significative entre r et 0 au risque $\alpha = 5\ \%$, on ne peut donc pas rejeter non plus H_0 au risque $\alpha = 1\ \%$.

2. Au seuil $\alpha = 1\ \%$, il faut que t soit supérieur à celui dans la table pour $\alpha = 0,02$:

$$t = \frac{r}{\sqrt{1-r^2}}\sqrt{n-2} > t_{2\%:n-2}$$

En remplacant par les valeurs connues, l'équation devient :

$$\frac{0,39}{\sqrt{1-0,39^2}}\sqrt{n-2} > t_{2\%:n-2} \Leftrightarrow 0,4235\sqrt{n-2} > t_{2\%:n-2}$$

On peut chercher la première valeur de n qui satisfait cette condition (en sachant d'après le problème précédent que n doit être supérieur à 18) :

$n = 18 \quad t_{2\%:n-2} = 1,746 \quad t = 0,4235\sqrt{18-2} = 1,694$

$n = 19 \quad t_{2\%:n-2} = 1,740 \quad t = 0,4235\sqrt{20-2} = 1,7463$

$n = 20 \quad t_{2\%:n-2} = 1,734 \quad t = 0,4235\sqrt{20-2} = 1,797$

On voit qu'à partir de n = 19, on aurait pu montrer une différence significative au seuil $\alpha = 1\ \%$.

GLOSSAIRE

Étendue (intervalle maximal ou amplitude)

L'étendue d'un caractère observé est la différence entre les valeurs extrêmes de ce caractère. Elle correspond à la valeur maximum des observations moins la valeur minimum de ces observations.

Groupes appariés - série de données appariées

On parle de groupes ou de séries de données appariés lorsque les données ne sont pas indépendantes, c'est-à-dire lorsque les échantillons sont dépendants, lorsque chaque sujet est son propre témoin. On parle aussi de données pairées.

Dans certains ouvrages, on rencontre le terme appareillé au lieu de apparié, cependant apparié est le terme statistique usuel.

Dans le cas de deux séries appariées, les données sont couplées, chaque sujet est comparé à lui-même. On dispose alors pour chaque sujet de deux observations, soit à deux temps différents (données répétées), soit selon deux méthodes différentes.

Dans le cas de plus de deux groupes appariés, les mêmes individus statistiques sont étudiés selon une ou plusieurs variables. Par exemple, si on étudie l'effet de trois thérapies différentes A, B et C, chez tous les sujets d'un même échantillon, on disposera de trois séries de données appariées à comparer. Ou encore, si on étudie l'effet d'une thérapie T, chez un seul et même groupe de sujets, quatre années de suite, on disposera de 4 séries de données appariées à comparer. Dans ce cas, on parle aussi de données répétées.

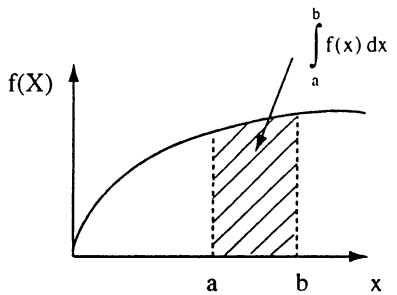
Groupes non appariés ou indépendants

Aux groupes appariés, on oppose les groupes non appariés ou indépen-

dants (ou non pairés). Chaque groupe est alors composé de sujets différents que l'on cherche à comparer selon les modalités de la ou des variables étudiées.

Intégrales

L'intégrale d'une fonction réelle f(x), notée $\int_a^b f(x)\,dx$, est l'aire sous la courbe sur l'intervalle [a, b] :



Médiane

La médiane correspond à la modalité centrale d'un caractère, séparant une population en deux parties égales. C'est la valeur qui partage l'ensemble des observations classées (dans l'ordre croissant ou décroissant) en deux moitiés de tailles égales, contenant donc chacune 50 % des observations. La médiane est la modalité correspondant à une fréquence cumulée de 0,5, ou à un pourcentage cumulé de 50 %.

Modalité

Le terme modalité correspond à une valeur donnée d'un caractère ou d'une variable qualitative.

Mode (ou dominante)

Le mode ou la dominante d'un caractère discontinu est la modalité qui a le plus fort effectif. Le mode d'un caractère continu est la modalité qui a la plus forte densité. C'est la valeur la plus probable d'un caractère. Dans le cas de caractères continus, le mode est sensible au choix du regroupement par classe.

Moyenne

La moyenne arithmétique ($\bar{x}$) d'un caractère quantitatif est le quotient de la somme de toutes les valeurs observées de ce caractère par le nombre total des observations.

$$\bar{x} = \frac{x_1 + \ldots + x_i + \ldots + x_n}{n} = \sum_{i=1}^{n} \frac{x_i}{n}$$

Primitives

La primitive d'une fonction réelle f(x) est la fonction de x qui a pour valeur l'aire sous la courbe sur l'intervalle [- ∞, x], x variant de - ∞ à + ∞ :

$\int_{-\infty}^{x} f(t)\,dt = F(x)$ avec t appelée variable d'intégration.

Si $F(x) = \int_{-\infty}^{x} f(t)\,dt$

alors $\frac{dF(x)}{dx} = f(x)$

Ainsi la primitive de f(x) est F(x) et la dérivée de F(x) est f(x).

Notation : $F(x) = \int f(x)\,dx$

Somme

Rappelons que l'on note la somme d'une suite de termes x_i :

$$\sum_{i=1}^{n} x_i = x_1 + x_2 + \ldots + x_n$$

et on peut mettre en facteur une constante et la sortir de la somme :

$\sum_{i=1}^{n} k\,x_i = k \sum_{i=1}^{n} x_i$ si k est une constante indépendante de i

Variable ou caractère

Chaque individu ou unité statistique, appartenant à une population, peut être décrit par un ensemble de caractéristiques appelées variables ou caractères. Ces variables peuvent être qualitatives ou quantitatives.

Variables qualitatives - variables quantitatives

Les variables quantitatives sont caractérisées par des quantités numériques : elles peuvent prendre des valeurs réelles sur un intervalle continu (variable quantitative continue) comme le temps, la taille, le poids ; ou constituer un intervalle d'entiers positifs (variable quantitative discrète) comme peut l'être l'âge compté en années sur l'intervalle [0 ; 130], le nombre d'enfants dans la famille, ou le quotient intellectuel.

Par opposition, les variables qualitatives sont définies par un ensemble de catégories, où chaque observation appartient à l'une des catégories (ou modalités). Les variables qualitatives peuvent être classées en variables nominales (catégorie socioprofessionnelle) ou en variables ordinales (niveau de vie classé bas, moyen, élevé).

Variable dépendante (variable à expliquer)
Variable indépendante (variable explicative)

En plus de la dichotomie variables qualitatives ou variables quantitatives, la statistique distingue les variables dépendantes (ou à expliquer) et les variables indépendantes (ou explicatives). Les variables dépendantes correspondent aux données, tandis que les variables indépendantes sont manipulées par l'expérimentateur. Par exemple, l'appartenance à un groupe (tel que celui des cas ou des témoins) est une variable indépendante, car il s'agit d'une variable contrôlée par l'expérimentateur. En effet, l'expérimentateur décide des traitements et du groupe de traitement pour chaque sujet étudié. Par ailleurs, l'expérimentateur peut être amené à contrôler les variables âge et sexe par exemple. Dans ce cas l'expérimentateur s'attachera à tirer au sort les sujets de façon à ce qu'ils soient représentatifs selon les variables âge et sexe de la population de référence. Ces variables indépendantes permettront alors éventuellement d'expliquer les données, c'est-à-dire les variables dépendantes (à expliquer) telles qu'un score d'anxiété, une note à un examen, un temps de réponse, un score de contrôle des événements...

INDEX

Index thématique détaillé

Analyse de variance (ANOVA), 26, 225-255
Analyse de variance à un facteur, 225-234, 240-247, 252-255
Analyse de variance à deux facteurs, 234-240, 247-250, 252-255
Analyse des données, 51
Analyse descriptive, 24
Appariés - non appariés, 93, 151, 317
Arbres de décision, 23
Arbres des probabilités, 74,76
Arrangements, 64-65
Bibliographiques (bases de données), 32-34
Cahier des charges, 43-44
Cahier d'observation, 44-46
Camemberts, 129-131
Caractère ou variable, 24
Cas continu, 79, 81, 84
Cas témoin, 93
Chi-deux (voir Khi-deux)
Choix des tests, 23-27
Codage et recodage des données, 50-51
Coefficient de corrélation, 19
Coefficient de variation, 90
Combinaisons, Combinatoire 63, 65
Comparaison de deux moyennes observées, 182-188, 204, 216
Comparaison de deux lois observées, 188-201

Comparaison entre deux variances, 201, 218
Comparaison entre une moyenne théorique et une moyenne observée, 175, 179
Comparaison entre plusieurs répartitions, 156-159
Comparaison entre plusieurs moyennes, 225
Conditions d'application, 26
Correction de Yates, 26, 159-162, 165
Corrélation, 26, 91, 288, 303-310
Corrélation linéaire, 288-289, 303-310
Covariance, 90-91
Degré de signification, 153, 177
Degrés de liberté, 153
Densité de probabilité, 81
Description de la population étudiée, 119-120
Description des données, 23, 119-124
Description des variables, 120, 121
Diagrammes en bâtons, 125-127
Distribution, 93
Distribution aléatoire, 93
Distribution normale, 95-103
Données catégorielles ou qualitatives, 24-27
Données manquantes, 37, 49-50
Données quantitatives, 24-27
Ecart-moyen, 86
Ecart-réduit, 176
Ecart-type, 19, 86-88, 176
Echantillon, 52,67
Echantillon apparié, 93
Echantillonnage, 52, 91-93
Echelles d'intervalles, numériques ou quantitatives, 48
Echelles de mesure, 47-48
Echelles nominales ou qualitatives, 47
Echelles ordinales, 48
Effectifs théoriques et observés, 152
Erreur d'échantillonnage, 67
Erreur de première espèce, 141
Erreur de seconde espèce, 142
Etendue (intervalle maximal ou amplitude), 86-87, 317
Evénement, 68-71
Evénements indépendants, 68-69
Exercices sur les probabilités et les intervalles, 110-116
Exercices sur le Khi-deux, 166-172

Exercices sur la comparaison de moyennes, 259-284
Exercices sur la corrélation et la régression, 311-315
Facteurs, 225
Faux positifs ou négatifs, 77
Fonction de répartition, 79
Groupes d'individus, 24
Glossaire, 317-320
Histogrammes, 127-129
Hypothèse, 31, 35
Hypothèse alternative, 135, 136
Hypothèse de recherche, 31
Hypothèse nulle, 135, 136
Indépendance, 73, 74, 317
Inférence statistique, 20
Inrégrales, 318
Internet (sites), 58-59
Interprétation des résultats, 52
Intervalle de confiance, 103
Intervalle de confiance d'un pourcentage, 109
Intervalle de confiance d'une corrélation, 307
Intervalle de confiance d'une moyenne, 107
Intervalle de pari d'un pourcentage, 108
Intervalle de pari d'une moyenne, 106
Kendall, 309
Khi-deux - Test du χ^2, 149-159, 165
Khi-deux corrigé de Yates, 159-162, 165
Khi-deux de Mac Nemar, 162-164, 165
Kruskall-Wallis - Test de Kruskall-Wallis, 27, 240-247, 254
Liaison entre deux variables qualitatives, 156
Liaison entre deux variables quantitatives, 287
Limite inférieure, 103
Limite supérieure, 103
Limites (intervalles) de confiance, 103
Littérature spécialisée, 32-34
Logiciels statistiques, 54-58
Loi binomiale, 93
Loi de Poisson, 94
Loi de probabilité - distribution de probabilité, 93
Loi normale, 67, 95-103
Mann et Whitney - Test de Mann et Whitney, 26, 188-201
Mc Nemar - Test du χ^2 de Mc Nemar, 26

Médiane, 81, 83, 318
Méthodologie, 31-59
Modalité, 46-47, 318
Modalités de réponses, 46-47, 318
Mode (ou dominante), 81, 84-85, 318
Moyenne, 19, 81-83, 318
Nombre de sujets nécessaires, 38-43
Normalité (voir loi normale)
Notation, 19
Paramètre, 19
Permutations, 63-64
Plan d'analyse statistique, 17, 22, 51
Population, 19, 91
Population théorique ou de référence, 135
Position, 81
Précision d'un pourcentage dans le cas de deux classes (voir intervalle de confiance)
Précision d'une moyenne (voir intervalle de confiance)
Prévalence, 77
Primitives, 319
Probabilité conditionnelle, 71-73
Probabilités, 63, 67
Procédure statistique résumée dans le cas de tests d'hypothèses, 52-54
Proportion, 19
Protocole, 35-37
Psychologie (Application en), 20-22
Puissance, 141-143
Questions, 46
Questionnaire, 44-48
Randomisation, 92
Récapitulatifs - *variables qualitatives, 165*
- *variables quantitatives dans le cas de 2 groupes, 216*
- *variables quantitatives dans le cas de plus de 2 groupes, 252*
Recherches bibliographiques, 32-34
Recueil des données, 48
Régression, 27, 288, 289-303, 309-310
Regroupement des données, 122
Répartition, 93
Répartition observée, 135
Répartition théorique, 135
Représentations des données, 125

Représentations graphiques, 125
Risque d'erreur, 104, 141-142
Saisie des données, 49
Sensibilité, 77
Seuil de rejet (voir zone de rejet)
Seuil de signification (voir degré de signification)
Spearman - ρ' de Spearman, 308-309
Statistique , 18
Statistiques, 19
Statistiques descriptives, 20, 119-131
Statistiques en psychologie, 20
Statistiques inférentielles, 20, 135-147
Somme, 319
Table de F, 336-343
Table de la loi normale (ε), 331
Table de Kolmogorov-Smirnov, 344-345
Table de Kruskal-Wallis, 349-350
Table de t, 332-333
Table du coefficient de corrélation ρ de Bravais Pearson, 351
Table du coefficient de corrélation ρ' de Spearman, 352
Table du test de Mann et Whitney, 346-347
Table du test de Wilcoxon, 348
Table du χ^2, 334-335
Tables statistiques, 330
Taille d'échantillon (voir conditions d'application et nombre de sujets)
Test bilatéral, 136, 176
Test de Bravais Pearson, 304-308
Test de Spearman, 308-309
Test d'indépendance entre deux variables quantitatives, 287
Test de corrélation des rangs, 27
Test de Kolmogorov-Smirnov, 144-147
Test de Kruskal-Wallis, 27, 240-247, 254
Test de Mann et Whitney, 26, 188-201, 216
Test de Wilcoxon, 26, 211-215, 220
Test du χ^2 corrigé de Yates, 26, 159-162, 165
Test du χ^2 de Mc Nemar, 26
Test t, 26, 216
Test t apparié, 26, 208-211, 216
Test t non apparié, 26, 179-182, 185-188, 220
Test unilatéral à droite ou à gauche, 138-140, 177-178
Test ε ou z (test de la loi normale), 26, 175, 182, 216

Test ε ou z apparié, 26, 204-208, 220
Test ε ou z non apparié, 26, 175-178, 182-185, 216
Tests d'hypothèses, 135-143
Tests non paramétriques, 26, 143
Tests paramétriques, 26, 143
Tests statistiques, 23
Théorème de Bayes, 76-79
Traitement des données, 48
Tri à plat des données, 123
Valeur prédictive (positive ou négative), 77
Variable aléatoire, 79
Variable dépendante - variable indépendante, 320
Variable expliquée - variable à expliquer, 320
Variable ou caractère, 24, 319
Variables continues ou discrètes, 79, 81, 84
Variables descriptives, 44
Variables qualitatives, 24-27, 149, 319
Variables quantitatives, 24-27, 173-221, 223-315, 319
Variance, 19, 88-90, 105, 201
Variance d'une moyenne, 104
Variance résiduelle, 227
Wilcoxon (voir test de Wilcoxon)
Yates, 159-162, 165
Zone de rejet, 95-103, 153

OUVRAGES DE RÉFÉRENCE

Armitage P., Berry G., *Statistical methods in medical research*, Ed. Blackwell Science, London, third edition, 1994.

Barbut M., *Mathématiques des sciences humaines*, Tome I et II, Presses Universitaires de France le psychologue, Paris, 1967 et 1968.

Bardin L., *L'analyse de contenu*, Presses Universitaires de France le psychologue, Paris, 1977.

Beaufils B., *Statistiques appliquées à la psychologie*, Tome 1 et 2, Ed. Bréal, 1996.

Cadet B., *Méthodes statistiques en psychologie*, Presses Universitaires de Caen, 1996.

Caulin C., Chastang C., Dahan R., *Méthodologie de l'évaluation thérapeutique*, Ed. Masson, Paris, 1993.

Cibois P., *L'analyse des données en sociologie*, Presses Universitaires de France, Paris, 1984.

Cibois P., *L'analyse factorielle*, Presses Universitaires de France, Paris, 1983.

Doise W., Clemence A., Lorenzi-Cioldi F., *Représentations sociales et analyses de données*, Presses Universitaires de Grenoble - vies sociales, 1992.

Ehrlich S., Flament C., *Précis de statistique*, Presses Universitaires de France le psychologue, Paris, 1961.

Ferréol G., Schlacther D., Rahmania N., Duverney D., *Dictionnaire des techniques quantitatives appliquées aux sciences économiques et sociales*, Ed. A. Colin, Paris, 1995.

Fénelon J-P., *Qu'est-ce que l'analyse des données ?*, Ed. Lefonen, Paris, 1981.

Gardner M.J., Altman D.G., *Statistics with confidence*, British Medical Journal, London, 1989.

Howell D.C., *Méthodes statistiques en sciences humaines*, Ed. De Boeck Université, Paris,1998.

Kendall M.G., Gibbons J.D., *Rank correlation methods*, Ed. Arnold, London, Fifth edition, 1990.

Langouet G., Porlier J-C., *Pratiques statistiques en sciences humaines et sociales*, ESF éditeur, Paris, 1989.

Lebart L., Morineau A., Fénelon J-P., *Traitement des données statistiques*, Ed. Dunod, Paris, 1979.

Lévy M. L., *Comprendre les statistiques*, Ed. du Seuil, 1979.

Matalon B., *L'analyse hierarchique*, Ed. Gauthier-Villars, Paris, 1965.

Mercier M., *Biostatistique et Probabilités - exercices, problèmes et épreuves corrigées*, Ed. Ellipses, Paris, 1996.

Mialaret G., *Statistiques*, Presses Universitaires de France, Paris, 1996.

Reuchlin M., *Introduction à la recherche en psychologie*, Ed. Nathan Université, 1992.

Robert C., *Analyse descriptive multivariée*, Ed. Médecine Sciences Flammarion, Paris, 1989.

Rossi J-P., *L'approche expérimentale en psychologie*, Ed. Dunod, Paris, 1997.

Rouanet H., Le Roux B., Bert M-C., *Statistique en sciences humaines : procédures naturelles*, Ed. Dunod, Paris, 1987.

Rouanet H., Le Roux B., *Exercices et solutions - Statistique en sciences humaines*, Ed. Dunod, Paris, 1995.

Rumeau-Rouquette G., Bréart G., Padieu R., *Méthodes en épidémiologie : Echantillonnage - Investigation - Analyse*, Ed. Flammarion Médecine Sciences, Paris, 1985.

Saporta G., *Probabilités, analyse des données et statistiques*, Ed. Technnip, Paris, 1990.

Schwartz D., *Le jeu de la science et du hasard*, Ed. Flammarion, Paris, 1994.

Schwartz D., *Méthodes statistiques à l'usage des médecins et des biologistes*, Ed. Flammarion Médecine Sciences, Paris, 1963.

Spiegel M.R., *Probabilités et statistique*, Série Schaum, 1981.

Spiegel M.R., *Théorie et applications de la statistique*, Série Schaum, 1972.

Tomassonne R., Lesquoy E., Miller C., *La régression : nouveaux regards sur une ancienne méthode statistique*, Ed. Masson, Paris, 1983.

Wonnacott T.H., Wonnacott R.J., *Statistique*, Ed. Economica, Paris, 1972.

TABLES STATISTIQUES

Les chiffres des tables statistiques ont été calculés par les auteurs de cet ouvrage à l'aide des logiciels MATHEMATICA® et EXCEL®.

Table de ε
Table de la loi Normale dans le cas d'un test bilatéral

Dans le cas bilatéral, la table donne la probabilité α pour que l'écart-réduit égale ou dépasse, en valeur absolue, une valeur donnée ε, c'est-à-dire la probabilité extérieure à l'intervalle [-ε ; +ε].
Autrement dit, la table de la loi normale donne la valeur de la probabilité α correspondant à la surface sous la courbe de la loi normale centrée réduite N(0,1) extérieure à l'intervalle [-ε ; +ε], soit $P(X > \varepsilon \text{ ou } X < -\varepsilon) = \alpha$. La probabilité complémentaire de la probabilité α correspond donc à la probabilité : $P(-\varepsilon < X < +\varepsilon) = 1-\alpha$.

Exemple : Le risque est α = 0,05 pour un écart-réduit ε = 1,960 (autrement dit, on a 5 % de chance d'obtenir en valeur absolue un écart-réduit calculé supérieur ou égal à 1,96)

α	ε	α	ε	α	ε
0,01	2,576	0,34	0,954	0,67	0,426
0,02	2,326	0,35	0,935	0,68	0,412
0,03	2,170	0,36	0,915	0,69	0,399
0,04	2,054	0,37	0,896	0,70	0,385
0,05	1,960	0,38	0,878	0,71	0,372
0,06	1,881	0,39	0,860	0,72	0,358
0,07	1,812	0,40	0,842	0,73	0,345
0,08	1,751	0,41	0,824	0,74	0,332
0,09	1,695	0,42	0,806	0,75	0,319
0,10	1,645	0,43	0,789	0,76	0,305
0,11	1,598	0,44	0,772	0,77	0,292
0,12	1,555	0,45	0,755	0,78	0,279
0,13	1,514	0,46	0,739	0,79	0,266
0,14	1,476	0,47	0,722	0,80	0,253
0,15	1,440	0,48	0,706	0,81	0,240
0,16	1,405	0,49	0,690	0,82	0,228
0,17	1,372	0,50	0,674	0,83	0,215
0,18	1,341	0,51	0,659	0,84	0,202
0,19	1,311	0,52	0,643	0,85	0,189
0,20	1,282	0,53	0,628	0,86	0,176
0,21	1,254	0,54	0,613	0,87	0,164
0,22	1,227	0,55	0,598	0,88	0,151
0,23	1,200	0,56	0,583	0,89	0,138
0,24	1,175	0,57	0,568	0,90	0,126
0,25	1,150	0,58	0,553	0,91	0,113
0,26	1,126	0,59	0,539	0,92	0,100
0,27	1,103	0,60	0,524	0,93	0,088
0,28	1,080	0,61	0,510	0,94	0,075
0,29	1,058	0,62	0,496	0,95	0,063
0,30	1,036	0,63	0,482	0,96	0,050
0,31	1,015	0,64	0,468	0,97	0,038
0,32	0,994	0,65	0,454	0,98	0,025
0,33	0,974	0,66	0,440	0,99	0,013

Extension de la table pour de faibles valeurs de epsilon :

α	idem	ε
0,1	1 E-01	1,645
0,01	1 E-02	2,576
0,001	1 E-03	3,290
...	1 E-04	3,891
	1 E-05	4,417
	1 E-06	4,892
	1 E-07	5,327
	1 E-08	5,731
	1 E-09	6,109

D'après Fisher et Yates, Statistical tables for biological, agricultural and medical research (Oliver and Boyd. Edinburgh)

Dans le cas unilatéral à droite comme à gauche, il suffit de diviser α par 2, et de lire dans cette même table.
Exemple : Le risque est α = 0,05 pour un écart-réduit ε = 1,645 (autrement dit, on a 5 % de chance d'obtenir un écart-réduit calculé supérieur ou égal à 1,645)

Table de t

La table donne la probabilité α pour que t égale ou dépasse une valeur de t calculée, en fonction du nombre de degrés de liberté (ddl).

ddl \ α	0,95	0,90	0,50	0,10	0,05	0,01	0,001
1	0,079	0,158	1,000	6,314	12,710	63,660	636,60
2	0,071	0,142	0,816	2,920	4,303	9,930	31,600
3	0,068	0,137	0,765	2,353	3,182	5,841	12,920
4	0,067	0,134	0,741	2,132	2,776	4,604	8,610
5	0,066	0,132	0,727	2,015	2,571	4,032	6,869
6	0,065	0,131	0,718	1,943	2,447	3,707	5,959
7	0,065	0,130	0,711	1,895	2,365	3,499	5,408
8	0,065	0,130	0,706	1,860	2,306	3,355	5,041
9	0,064	0,129	0,703	1,833	2,262	3,250	4,781
10	0,064	0,129	0,700	1,812	2,228	3,169	4,587
11	0,064	0,129	0,697	1,796	2,201	3,106	4,437
12	0,064	0,128	0,695	1,782	2,179	3,055	4,318
13	0,064	0,128	0,694	1,771	2,160	3,012	4,221
14	0,064	0,128	0,692	1,761	2,145	2,977	4,140
15	0,064	0,128	0,691	1,753	2,131	2,947	4,073
16	0,064	0,128	0,690	1,746	2,120	2,921	4,015
17	0,064	0,128	0,689	1,740	2,110	2,898	3,965
18	0,064	0,127	0,688	1,734	2,101	2,878	3,922
19	0,064	0,127	0,688	1,729	2,093	2,861	3,883
20	0,063	0,127	0,687	1,725	2,086	2,845	3,850
21	0,063	0,127	0,686	1,721	2,080	2,831	3,819
22	0,063	0,127	0,686	1,717	2,074	2,819	3,792
23	0,063	0,127	0,685	1,714	2,069	2,807	3,768
24	0,063	0,127	0,685	1,711	2,064	2,797	3,745
25	0,063	0,127	0,684	1,708	2,060	2,787	3,725
26	0,063	0,127	0,684	1,706	2,056	2,779	3,707
27	0,063	0,127	0,684	1,703	2,052	2,771	3,689
28	0,063	0,127	0,683	1,701	2,048	2,763	3,674
29	0,063	0,127	0,683	1,699	2,045	2,756	3,660
30	0,063	0,127	0,683	1,697	2,042	2,750	3,646
31	0,063	0,127	0,682	1,696	2,040	2,744	3,633
32	0,063	0,127	0,682	1,694	2,037	2,738	3,622
33	0,063	0,127	0,682	1,692	2,035	2,733	3,611
34	0,063	0,127	0,682	1,691	2,032	2,728	3,601
35	0,063	0,127	0,682	1,690	2,030	2,724	3,591
36	0,063	0,127	0,681	1,688	2,028	2,719	3,582
37	0,063	0,127	0,681	1,687	2,026	2,715	3,574
38	0,063	0,127	0,681	1,686	2,024	2,712	3,566
39	0,063	0,126	0,681	1,685	2,023	2,708	3,558
40	0,063	0,126	0,681	1,684	2,021	2,704	3,551
41	0,063	0,126	0,681	1,683	2,020	2,701	3,544
42	0,063	0,126	0,680	1,682	2,018	2,698	3,538
43	0,063	0,126	0,680	1,681	2,017	2,695	3,532
44	0,063	0,126	0,680	1,680	2,015	2,692	3,526
45	0,063	0,126	0,680	1,679	2,014	2,690	3,520
46	0,063	0,126	0,680	1,679	2,013	2,687	3,515
47	0,063	0,126	0,680	1,678	2,012	2,685	3,510
48	0,063	0,126	0,680	1,677	2,011	2,682	3,505
49	0,063	0,126	0,680	1,677	2,010	2,680	3,500

Exemple : avec ddl = 5, le risque $\alpha = 0,05$ pour un $t_T = 2,571$ (autrement dit, on a 5 % de chance d'obtenir un t calculé supérieur ou égal à 2,571).

D'après Fisher et Yates, Statistical tables for biological, agricultural and medical research, Ed. Oliver and Boyd, Edinburgh (Great Britain).

Table de t (suite)

ddl \ α	0,95	0,90	0,50	0,10	0,05	0,01	0,001
50	0.0630	0.126	0.679	1.676	2,009	2.678	3.496
51	0,0630	0,126	0.679	1.675	2,008	2.676	3.492
52	0.0630	0,126	0.679	1.675	2,007	2.674	3.488
53	0,0630	0,126	0.679	1.674	2,006	2.672	3.484
54	0.0630	0,126	0.679	1.674	2,005	2.670	3.480
55	0.0630	0,126	0.679	1.673	2,004	2.668	3.476
56	0.0630	0,126	0.679	1.673	2,003	2.667	3.473
57	0.0630	0.126	0.679	1.672	2,002	2.665	3.469
58	0.0630	0,126	0.679	1.672	2.002	2.663	3.466
59	0.0630	0.126	0.679	1.671	2,001	2.662	3,463
60	0,0630	0,126	0.679	1.671	2,000	2.660	3,460
61	0.0630	0,126	0.679	1.670	2.000	2.659	3.457
62	0.0630	0,126	0.678	1.670	1,999	2.657	3.454
63	0.0630	0.126	0.678	1.669	1,998	2.656	3.452
64	0.0630	0.126	0.678	1.669	1,998	2.655	3.449
65	0.0630	0.126	0.678	1.669	1.997	2.654	3.447
66	0,0630	0.126	0,678	1.668	1,997	2.652	3,444
67	0,0630	0,126	0.678	1.668	1,996	2.651	3.442
68	0,0630	0.126	0.678	1.668	1,995	2.650	3,439
69	0.0630	0,126	0.678	1.667	1,995	2.649	3.437
70	0.0630	0.126	0.678	1.667	1,994	2.648	3,435
71	0,0630	0,126	0,678	1.667	1,994	2,647	3.433
72	0,0630	0,126	0,678	1.666	1,993	2,646	3,431
73	0,0630	0,126	0,678	1.666	1,993	2.645	3,429
74	0.0630	0,126	0.678	1.666	1,993	2.644	3,427
75	0,0630	0.126	0.678	1.665	1,992	2.643	3.425
76	0.0630	0.126	0,678	1.665	1,992	2.642	3,423
77	0.0630	0.126	0.678	1.665	1.991	2.641	3.421
78	0.0630	0.126	0.678	1.665	1,991	2.640	3.420
79	0.0630	0.126	0.678	1.664	1,990	2.639	3.418
80	0.0630	0.126	0.678	1.664	1,990	2.639	3.416
81	0.0630	0,126	0.678	1.664	1.990	2.638	3.415
82	0,0630	0,126	0.677	1.664	1,989	2.637	3,413
83	0,0630	0,126	0.677	1.663	1,989	2.636	3,412
84	0.0630	0,126	0.677	1.663	1,989	2.636	3.410
85	0.0630	0.126	0.677	1.663	1,988	2.635	3.409
86	0.0630	0.126	0.677	1.663	1,988	2.634	3.407
87	0.0630	0.126	0.677	1.663	1,988	2.634	3,406
88	0.0630	0,126	0.677	1.662	1,987	2.633	3,405
89	0.0630	0,126	0.677	1.662	1,987	2,632	3,403
90	0.0630	0,126	0.677	1.662	1,987	2.632	3.402
91	0.0630	0,126	0.677	1.662	1,986	2.631	3.401
92	0.0630	0,126	0.677	1.662	1.986	2.630	3.399
93	0.0630	0.126	0.677	1.661	1.986	2.630	3.398
94	0.0630	0.126	0.677	1.661	1.986	2.629	3.397
95	0.0630	0.126	0.677	1.661	1.985	2.629	3.396
96	0.0630	0.126	0.677	1.661	1.985	2.628	3.395
97	0.0630	0.126	0.677	1.661	1.985	2.627	3.394
98	0,0630	0.126	0.677	1.661	1,984	2.627	3,393
99	0,0630	0.126	0.677	1.660	1.984	2.626	3.391
100	0.0630	0.126	0.677	1.660	1,984	2.626	3.390
150	0.0630	0,126	0.676	1.655	1.976	2.609	3.357
200	0.0630	0,126	0.676	1.653	1,972	2.601	3.340
∞	0.0630	0,126	0.675	1.646	1,962	2.581	3.300

Table du χ^2

La table donne la probabilité α pour que le χ^2 égale ou dépasse une valeur de χ^2 calculée, en fonction du nombre de degrés de liberté (ddl).

ddl \ α	0,95	0,90	0,50	0,10	0,05	0,02	0,01	0,002	0,001
1	0,004	0,016	0,455	2,706	3,841	5,412	6,635	9,549	10,830
2	0,103	0,211	1,386	4,605	5,991	7,824	9,210	12,429	13,820
3	0,352	0,584	2,366	6,251	7,815	9,837	11,345	14,796	16,270
4	0,711	1,064	3,357	7,779	9,488	11,668	13,277	16,923	18,470
5	1,145	1,610	4,351	9,236	11,070	13,388	15,086	18,908	20,510
6	1,635	2,204	5,348	10,645	12,592	15,033	16,812	20,791	22,460
7	2,167	2,833	6,346	12,017	14,067	16,622	18,475	22,601	24,320
8	2,733	3,490	7,344	13,362	15,507	18,168	20,090	24,352	26,120
9	3,325	4,168	8,343	14,684	16,919	19,679	21,666	26,056	27,880
10	3,940	4,865	9,342	15,987	18,307	21,161	23,209	27,721	29,590
11	4,575	5,578	10,341	17,275	19,675	22,618	24,725	29,354	31,260
12	5,226	6,304	11,340	18,549	21,026	24,054	26,217	30,957	32,910
13	5,892	7,041	12,340	19,812	22,362	25,471	27,688	32,536	34,530
14	6,571	7,790	13,339	21,064	23,685	26,873	29,141	34,091	36,120
15	7,261	8,547	14,339	22,307	24,996	28,259	30,578	35,627	37,700
16	7,962	9,312	15,338	23,542	26,296	29,633	32,000	37,146	39,250
17	8,672	10,085	16,338	24,769	27,587	30,995	33,409	38,648	40,790
18	9,390	10,865	17,338	25,989	28,869	32,346	34,805	40,136	42,310
19	10,117	11,651	18,338	27,204	30,144	33,687	36,191	41,610	43,820
20	10,851	12,443	19,337	28,412	31,410	35,020	37,566	43,072	45,310
21	11,591	13,240	20,337	29,615	32,671	36,343	38,932	44,522	46,800
22	12,338	14,041	21,337	30,813	33,924	37,659	40,289	45,961	48,270
23	13,091	14,848	22,337	32,007	35,172	38,968	41,638	47,392	49,730
24	13,848	15,659	23,337	33,196	36,415	40,270	42,980	48,811	51,180
25	14,611	16,473	24,337	34,382	37,652	41,566	44,314	50,223	52,620
26	15,379	17,292	25,336	35,563	38,885	42,856	45,642	51,627	54,050
27	16,151	18,114	26,336	36,741	40,113	44,140	46,963	53,022	55,480
28	16,928	18,939	27,336	37,916	41,337	45,419	48,278	54,411	56,890
29	17,708	19,768	28,336	39,087	42,557	46,693	49,588	55,792	58,300
30	18,493	20,599	29,336	40,256	43,773	47,962	50,892	57,167	59,700
31	19,281	21,434	30,336	41,422	44,985	49,226	52,191	58,536	61,100
32	20,072	22,271	31,336	42,585	46,194	50,487	53,486	59,899	62,490
33	20,867	23,110	32,336	43,745	47,400	51,743	54,775	61,256	63,870
34	21,664	23,952	33,336	44,903	48,602	52,995	56,061	62,607	65,250
35	22,465	24,797	34,336	46,059	49,802	54,244	57,342	63,955	66,620
36	23,269	25,643	35,336	47,212	50,998	55,489	58,619	65,296	67,980
37	24,075	26,492	36,336	48,363	52,192	56,730	59,893	66,634	69,350
38	24,884	27,343	37,335	49,513	53,384	57,969	61,162	67,966	70,700
39	25,695	28,196	38,335	50,660	54,572	59,204	62,428	69,293	72,060
40	26,509	29,051	39,335	51,805	55,758	60,436	63,691	70,617	73,400
41	27,326	29,907	40,335	52,949	56,942	61,665	64,950	71,938	74,740
42	28,144	30,765	41,335	54,090	58,124	62,892	66,206	73,253	76,080
43	28,965	31,625	42,335	55,230	59,304	64,116	67,459	74,566	77,420
44	29,787	32,487	43,335	56,369	60,481	65,337	68,710	75,874	78,750
45	30,612	33,350	44,335	57,505	61,656	66,555	69,957	77,179	80,080
46	31,439	34,215	45,335	58,641	62,830	67,771	71,201	78,481	81,400
47	32,268	35,081	46,335	59,774	64,001	68,985	72,443	79,779	82,720
48	33,098	35,949	47,335	60,907	65,171	70,197	73,683	81,075	84,040
49	33,930	36,818	48,335	62,038	66,339	71,406	74,919	82,367	85,350

d'après Fisher et Yates, Statistical tables for biological, agricultural and medical research, Ed. Oliver and Boyd, Edinburgh (Great Britain).

Table du χ^2 (suite)

ddl \ α	0,95	0,90	0,50	0,10	0,05	0,02	0,01	0,002	0,001
50	34,764	37.689	49,335	63,167	67,505	72,613	76,154	83.656	86.660
51	35.600	38,560	50,335	64.295	68,669	73,818	77,386	84.944	87,967
52	36.437	39,433	51,335	65.422	69,832	75,021	78,616	86,227	89,272
53	37.276	40,308	52,335	66,548	70,993	76,223	79,843	87,508	90,573
54	38.116	41.183	53,335	67,673	72,153	77,422	81,069	88,785	91.871
55	38,958	42,060	54,335	68,796	73,311	78.619	82.292	90.061	93,167
56	39,801	42,937	55,335	69,919	74,468	79,815	83,514	91,335	94.462
57	40,646	43,816	56,335	71,040	75,624	81,009	84,733	92,605	95,750
58	41,492	44,696	57,335	72,160	76,778	82,201	85,950	93,873	97.038
59	42,339	45,577	58,335	73,279	77,930	83,391	87,166	95,140	98.324
60	43,188	46.459	59,335	74,397	79,082	84,580	88,379	96.403	99,608
61	44,038	47,342	60,335	75,514	80,232	85,767	89.591	97,665	100.887
62	44,889	48.226	61,335	76,630	81,381	86,953	90,802	98,924	102.165
63	45,741	49,111	62,335	77.745	82,529	88,137	92.010	100,182	103.442
64	46,595	49.996	63,335	78,860	83,675	89,320	93,217	101,437	104.717
65	47,450	50.883	64.335	79,973	84.821	90.501	94.422	102.691	105.988
66	48.305	51.770	65,335	81,085	85,965	91,681	95,626	103.943	107.257
67	49.162	52.659	66,335	82,197	87,108	92,860	96.828	105.192	108.525
68	50,020	53,548	67,335	83,308	88,250	94,037	98,028	106,439	109.793
69	50,879	54.438	68,334	84,418	89,391	95,213	99.227	107.686	111.055
70	51,739	55,329	69,334	85,527	90,531	96,387	100.425	108.930	112.317
71	52,600	56.221	70,334	86,635	91,670	97,561	101.621	110.172	113.577
72	53,462	57.113	71,334	87.743	92,808	98.733	102.816	111,412	114.834
73	54,325	58.006	72,334	88,850	93,945	99,904	104.010	112,652	116.092
74	55.189	58,900	73,334	89,956	95,081	101.074	105.202	113,889	117.347
75	56.054	59.795	74,334	91,061	96,217	102,243	106,393	115,124	118.599
76	56,920	60.690	75,334	92.166	97,351	103,410	107,582	116,359	119.850
77	57,786	61,586	76,334	93,270	98,484	104,576	108,771	117,592	121,101
78	58.654	62,483	77,334	94,374	99,617	105,742	109,958	118,822	122,347
79	59,522	63,380	78,334	95,476	100,749	106,906	111,144	120.052	123.595
80	60,391	64.278	79,334	96.578	101.879	108.069	112.329	121.280	124.839
81	61.262	65.176	80,334	97.680	103,010	109.231	113,512	122.507	126.084
82	62,132	66.076	81,334	98,780	104,139	110.393	114,695	123,732	127.324
83	63.004	66.976	82,334	99,880	105,267	111.553	115.876	124,957	128,565
84	63,876	67,876	83,334	100.980	106,395	112,712	117.057	126,179	129.802
85	64,749	68,777	84,334	102,079	107,522	113,871	118,236	127,401	131.043
86	65,623	69.679	85,334	103.177	108,648	115.028	119,414	128.621	132.276
87	66,498	70,581	86,334	104.275	109,773	116.184	120.591	129,839	133,511
88	67,373	71.484	87,334	105,372	110,898	117,340	121,767	131.056	134.746
89	68,249	72,387	88,334	106,469	112,022	118,495	122,942	132,273	135,977
90	69,126	73,291	89,334	107,565	113,145	119.648	124,116	133,489	137,208
91	70,003	74,196	90,334	108,661	114,268	120,801	125,289	134,703	138.437
92	70,882	75,100	91,334	109,756	115,390	121,953	126,462	135,915	139.667
93	71,760	76,006	92,334	110,850	116,511	123,105	127.633	137,126	140.894
94	72,640	76,912	93,334	111,944	117.632	124,255	128,803	138,337	142,118
95	73,520	77.818	94,334	113,038	118,752	125.405	129,973	139.546	143.343
96	74.401	78.725	95,334	114.131	119.871	126.554	131.141	140.755	144.566
97	75,282	79.633	96,334	115,223	120.990	127.702	132,309	141,961	145.789
98	76.164	80,541	97,334	116,315	122.108	128.849	133.476	143.167	147.009
99	77.046	81,449	98,334	117.407	123,225	129,996	134,641	144,373	148.230
100	77.929	82,358	99,334	118,498	124,342	131,142	135,807	145.576	149.449
150	122.692	128,275	149,334	172.581	179,581	187.678	193,207	204,722	209.265
200	168,279	174,835	199,334	226.021	233.994	243,187	249.445	262.434	267.539

Nota Bene : pour de grandes valeurs de ddl, vous pouvez utiliser la table de la loi normale en faisant la transformation $\sqrt{2\chi^2}$ à comparer à la valeur seuil pour un ddl = $\sqrt{2\text{ddl} - 1}$ avec une variance de 1.

Table de F (seuil 5 %)

La table donne la limite supérieure de $F_B^A = S_A^2 / S_B^2$, pour le risque 5 % (valeur ayant 5 chances sur 100 d'être égalée ou dépassée), en fonction du nombre de degrés de liberté, ddl_A et ddl_B

Exemple : avec $ddl_A = 10$ et $ddl_B = 70$, la limite supérieure de F est $F_{70}^{10} = 1,97$

ddl(B) \ ddl(A)	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
1	161,45	199,50	215,71	224,58	230,16	233,99	236,77	238,88	240,54	241,88	242,98	243,91	244,69	245,36	245,95	246,47	246,92
2	18,513	19,000	19,164	19,247	19,296	19,329	19,353	19,371	19,385	19,396	19,405	19,412	19,419	19,424	19,429	19,433	19,437
3	10,128	9,552	9,277	9,117	9,013	8,941	8,887	8,845	8,812	8,785	8,763	8,745	8,729	8,715	8,703	8,692	8,683
4	7,709	6,944	6,591	6,388	6,256	6,163	6,094	6,041	5,999	5,964	5,936	5,912	5,891	5,873	5,858	5,844	5,832
5	6,608	5,786	5,409	5,192	5,050	4,950	4,876	4,818	4,772	4,735	4,704	4,678	4,655	4,636	4,619	4,604	4,590
6	5,987	5,143	4,757	4,534	4,387	4,284	4,207	4,147	4,099	4,060	4,027	4,000	3,976	3,956	3,938	3,922	3,908
7	5,591	4,737	4,347	4,120	3,972	3,866	3,787	3,726	3,677	3,637	3,603	3,575	3,550	3,529	3,511	3,494	3,480
8	5,318	4,459	4,066	3,838	3,688	3,581	3,500	3,438	3,388	3,347	3,313	3,284	3,259	3,237	3,218	3,202	3,187
9	5,117	4,256	3,863	3,633	3,482	3,374	3,293	3,230	3,179	3,137	3,102	3,073	3,048	3,025	3,006	2,989	2,974
10	4,965	4,103	3,708	3,478	3,326	3,217	3,135	3,072	3,020	2,978	2,943	2,913	2,887	2,865	2,845	2,828	2,812
11	4,844	3,982	3,587	3,357	3,204	3,095	3,012	2,948	2,896	2,854	2,818	2,788	2,761	2,739	2,719	2,701	2,685
12	4,747	3,885	3,490	3,259	3,106	2,996	2,913	2,849	2,796	2,753	2,717	2,687	2,660	2,637	2,617	2,599	2,583
13	4,667	3,806	3,411	3,179	3,025	2,915	2,832	2,767	2,714	2,671	2,635	2,604	2,577	2,554	2,533	2,515	2,499
14	4,600	3,739	3,344	3,112	2,958	2,848	2,764	2,699	2,646	2,602	2,565	2,534	2,507	2,484	2,463	2,445	2,428
15	4,543	3,682	3,287	3,056	2,901	2,790	2,707	2,641	2,588	2,544	2,507	2,475	2,448	2,424	2,403	2,385	2,368
16	4,494	3,634	3,239	3,007	2,852	2,741	2,657	2,591	2,538	2,494	2,456	2,425	2,397	2,373	2,352	2,333	2,317
17	4,451	3,592	3,197	2,965	2,810	2,699	2,614	2,548	2,494	2,450	2,413	2,381	2,353	2,329	2,308	2,289	2,273
18	4,414	3,555	3,160	2,928	2,773	2,661	2,577	2,510	2,456	2,412	2,374	2,342	2,314	2,290	2,269	2,250	2,233
19	4,381	3,522	3,127	2,895	2,740	2,628	2,544	2,477	2,423	2,378	2,340	2,308	2,280	2,256	2,234	2,215	2,198
20	4,351	3,493	3,098	2,866	2,711	2,599	2,514	2,447	2,393	2,348	2,310	2,278	2,250	2,225	2,203	2,184	2,167
25	4,242	3,385	2,991	2,759	2,603	2,490	2,405	2,337	2,282	2,236	2,198	2,165	2,136	2,111	2,089	2,069	2,051
30	4,171	3,316	2,922	2,690	2,534	2,421	2,334	2,266	2,211	2,165	2,126	2,092	2,063	2,037	2,015	1,995	1,976
35	4,121	3,267	2,874	2,641	2,485	2,372	2,285	2,217	2,161	2,114	2,075	2,041	2,012	1,986	1,963	1,942	1,924
40	4,085	3,232	2,839	2,606	2,449	2,336	2,249	2,180	2,124	2,077	2,038	2,003	1,974	1,948	1,924	1,904	1,885
45	4,057	3,204	2,812	2,579	2,422	2,308	2,221	2,152	2,096	2,049	2,009	1,974	1,945	1,918	1,895	1,874	1,855
50	4,034	3,183	2,790	2,557	2,400	2,286	2,199	2,130	2,073	2,026	1,986	1,952	1,921	1,895	1,871	1,850	1,831
60	4,001	3,150	2,758	2,525	2,368	2,254	2,167	2,097	2,040	1,993	1,952	1,917	1,887	1,860	1,836	1,815	1,796
70	3,978	3,128	2,736	2,503	2,346	2,231	2,143	2,074	2,017	1,969	1,928	1,893	1,863	1,836	1,812	1,790	1,771
80	3,960	3,111	2,719	2,486	2,329	2,214	2,126	2,056	1,999	1,951	1,910	1,875	1,845	1,817	1,793	1,772	1,752
90	3,947	3,098	2,706	2,473	2,316	2,201	2,113	2,043	1,986	1,938	1,897	1,861	1,830	1,803	1,779	1,757	1,737
100	3,936	3,087	2,696	2,463	2,305	2,191	2,103	2,032	1,975	1,927	1,886	1,850	1,819	1,792	1,768	1,746	1,726
∞	3,841	2,996	2,605	2,372	2,214	2,099	2,010	1,938	1,880	1,831	1,789	1,752	1,720	1,692	1,666	1,644	1,623

d'après E.F. Pearson et H.O. Hartley, Bio Metrika tables for statisticans, vol. 1 University Press, Cambridge.

Table de F (seuil 5 %) [suite]

ddl(B) \ ddl(A)	18	19	20	25	30	35	40	45	50	60	70	80	90	100	∞
1	247.32	247.69	248.02	249,26	250,10	250.69	251.14	251.49	251,77	252,20	252,50	252,72	252.90	253.04	254,32
2	19.440	19.443	19.446	19,456	19,463	19,467	19,471	19.473	19,476	19,479	19.481	19.483	19,485	19,486	19,496
3	8.675	8.667	8,660	8,634	8.617	8,604	8.594	8.587	8,581	8.572	8.566	8.561	8,557	8.554	8.526
4	5.821	5.811	5,803	5.769	5,746	5,729	5.717	5.707	5.699	5.688	5.679	5.673	5.668	5.664	5.628
5	4.579	4.568	4.558	4.521	4.496	4.478	4.464	4.453	4.444	4.431	4.422	4.415	4.409	4.405	4.365
6	3.896	3.884	3,874	3.835	3.808	3.789	3.774	3.763	3.754	3.740	3.730	3.722	3.716	3.712	3.669
7	3,467	3.455	3,445	3.404	3,376	3,356	3.340	3.328	3.319	3.304	3.294	3,286	3.280	3.275	3.230
8	3.173	3.161	3,150	3.108	3,079	3.059	3.043	3,030	3.020	3,005	2.994	2.986	2.980	2.975	2.928
9	2.960	2,948	2,936	2.893	2,864	2.842	2.826	2.813	2,803	2,787	2.776	2.768	2,761	2.756	2.707
10	2.798	2.785	2.774	2,730	2.700	2,678	2.661	2.648	2.637	2.621	2.609	2.601	2.594	2.588	2.538
11	2.671	2,658	2,646	2.601	2,570	2,548	2,531	2.517	2.507	2.490	2.478	2,469	2.462	2.457	2.404
12	2.568	2.555	2,544	2,498	2,466	2,443	2.426	2,412	2.401	2,384	2,372	2.363	2,356	2,350	2.296
13	2.484	2,471	2,459	2,412	2,380	2,357	2,339	2.325	2,314	2.297	2.284	2.275	2.267	2,261	2.206
14	2,413	2,400	2,388	2,341	2.308	2,284	2,266	2,252	2,241	2,223	2,210	2,201	2.193	2.187	2,131
15	2,353	2.340	2,328	2,280	2.247	2,223	2.204	2.190	2.178	2.160	2.147	2.137	2.130	2.123	2,066
16	2.302	2.288	2,276	2.227	2,194	2.169	2.151	2.136	2.124	2,106	2.093	2.083	2.075	2.068	2.010
17	2.257	2.243	2.230	2,181	2.148	2.123	2,104	2.089	2,077	2.058	2.045	2.035	2.027	2.020	1.960
18	2.217	2.203	2,191	2,141	2.107	2.082	2.063	2.048	2.035	2.017	2.003	1.993	1.985	1.978	1.917
19	2.182	2.168	2,155	2.106	2.071	2.046	2.026	2.011	1.999	1.980	1.966	1.955	1.947	1.940	1.878
20	2.151	2.137	2,124	2.074	2.039	2.013	1.994	1.978	1.966	1.946	1.932	1.922	1.913	1.907	1.843
25	2.035	2.021	2,007	1,955	1,919	1,892	1.872	1.855	1.842	1.822	1.807	1.796	1.787	1.779	1.711
30	1.960	1.945	1,932	1.878	1.841	1.813	1.792	1.775	1.761	1.740	1.724	1.712	1.703	1.695	1.622
35	1.907	1.892	1,878	1.824	1,786	1,757	1,735	1.718	1.703	1.681	1.665	1.652	1.643	1.635	1.558
40	1.868	1.853	1.839	1,783	1.744	1,715	1.693	1.675	1.660	1.637	1.621	1.608	1,597	1.589	1.509
45	1.838	1.823	1,808	1,752	1,713	1.683	1.660	1,642	1,626	1.603	1.586	1.573	1.562	1.554	1.470
50	1.814	1.798	1,784	1,727	1,687	1.657	1.634	1.615	1.599	1.576	1.558	1.544	1.534	1.525	1.438
60	1,778	1.763	1.748	1,690	1,649	1.618	1.594	1.575	1.559	1.534	1.516	1.502	1.491	1.481	1,389
70	1.753	1.737	1,722	1.664	1,622	1,591	1.566	1.546	1,530	1.505	1.486	1.471	1.459	1.450	1.353
80	1.734	1.718	1,703	1,644	1.602	1.570	1.545	1.525	1.508	1.482	1.463	1.448	1.436	1.426	1.325
90	1.720	1.703	1.688	1,629	1,586	1.554	1.528	1.508	1.491	1.465	1.445	1.429	1.417	1.407	1.302
100	1.708	1.691	1.676	1.616	1.573	1.541	1.515	1.494	1.477	1.450	1.430	1.415	1.402	1.392	1.283
∞	1.604	1.587	1.571	1.506	1.459	1.423	1.394	1.370	1.350	1.318	1.293	1.274	1.257	1.243	1.000

Table de F (seuil 2 %)

ddl(B) \ ddl(A)	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
1	1012,5	1249,5	1350,5	1405,8	1440,6	1464,4	1481,8	1495,0	1505,4	1513,7	1520,6	1526,3	1531,2	1535,4	1539,0	1542,3	1545,1
2	48,505	49,000	49,165	49,249	49,298	49,332	49,356	49,373	49,387	49,398	49,407	49,415	49,422	49,427	49,431	49,436	49,440
3	20,618	18,858	18,110	17,694	17,429	17,245	17,110	17,007	16,926	16,860	16,805	16,759	16,720	16,686	16,657	16,631	16,609
4	14,040	12,142	11,343	10,899	10,616	10,419	10,274	10,162	10,074	10,003	9,944	9,894	9,851	9,815	9,783	9,755	9,730
5	11,323	9,454	8,670	8,233	7,953	7,758	7,614	7,503	7,415	7,344	7,285	7,235	7,192	7,156	7,123	7,095	7,070
6	9,876	8,052	7,287	6,859	6,585	6,393	6,251	6,141	6,055	5,984	5,925	5,876	5,833	5,797	5,765	5,737	5,712
7	8,988	7,203	6,454	6,035	5,765	5,576	5,435	5,327	5,241	5,171	5,113	5,064	5,021	4,985	4,953	4,925	4,900
8	8,390	6,637	5,901	5,489	5,223	5,036	4,897	4,790	4,705	4,635	4,577	4,528	4,486	4,449	4,417	4,389	4,364
9	7,960	6,234	5,510	5,103	4,839	4,655	4,517	4,410	4,325	4,256	4,198	4,149	4,107	4,071	4,039	4,011	3,986
10	7,638	5,934	5,218	4,816	4,555	4,371	4,235	4,129	4,044	3,975	3,917	3,868	3,826	3,790	3,758	3,730	3,705
11	7,388	5,701	4,993	4,594	4,336	4,153	4,017	3,912	3,828	3,758	3,701	3,652	3,610	3,573	3,542	3,513	3,488
12	7,188	5,516	4,814	4,419	4,162	3,980	3,845	3,740	3,656	3,587	3,529	3,480	3,438	3,402	3,370	3,341	3,316
13	7,024	5,366	4,669	4,276	4,020	3,840	3,705	3,600	3,516	3,447	3,390	3,341	3,299	3,262	3,230	3,201	3,176
14	6,888	5,241	4,549	4,158	3,904	3,724	3,589	3,485	3,401	3,332	3,274	3,225	3,183	3,146	3,114	3,086	3,060
15	6,773	5,135	4,447	4,058	3,805	3,626	3,492	3,387	3,303	3,235	3,177	3,128	3,086	3,049	3,017	2,988	2,963
16	6,674	5,046	4,361	3,974	3,721	3,543	3,409	3,304	3,221	3,152	3,094	3,045	3,003	2,966	2,934	2,905	2,879
17	6,589	4,968	4,286	3,901	3,649	3,471	3,337	3,233	3,149	3,080	3,023	2,973	2,931	2,894	2,862	2,833	2,807
18	6,515	4,900	4,221	3,837	3,586	3,408	3,275	3,171	3,087	3,018	2,960	2,911	2,869	2,832	2,799	2,770	2,745
19	6,449	4,840	4,164	3,781	3,531	3,353	3,220	3,116	3,032	2,963	2,906	2,856	2,814	2,777	2,744	2,715	2,690
20	6,391	4,788	4,113	3,731	3,482	3,304	3,171	3,067	2,984	2,915	2,857	2,808	2,765	2,728	2,695	2,666	2,641
25	6,176	4,593	3,928	3,549	3,302	3,126	2,993	2,890	2,806	2,737	2,679	2,629	2,587	2,549	2,516	2,487	2,461
30	6,038	4,470	3,809	3,434	3,188	3,012	2,880	2,777	2,693	2,624	2,566	2,516	2,473	2,435	2,402	2,372	2,346
35	5,942	4,384	3,727	3,354	3,109	2,934	2,802	2,699	2,615	2,546	2,487	2,437	2,394	2,356	2,323	2,293	2,266
40	5,872	4,321	3,667	3,295	3,051	2,877	2,745	2,641	2,558	2,488	2,430	2,380	2,336	2,298	2,264	2,235	2,208
45	5,818	4,273	3,622	3,251	3,007	2,833	2,701	2,598	2,514	2,444	2,386	2,336	2,292	2,254	2,220	2,190	2,163
50	5,776	4,235	3,585	3,215	2,972	2,798	2,667	2,563	2,479	2,410	2,351	2,301	2,257	2,219	2,185	2,154	2,127
60	5,713	4,179	3,532	3,163	2,921	2,747	2,616	2,512	2,428	2,359	2,300	2,249	2,205	2,167	2,133	2,102	2,075
70	5,668	4,139	3,494	3,127	2,885	2,711	2,580	2,476	2,392	2,323	2,264	2,213	2,169	2,130	2,096	2,065	2,038
80	5,635	4,110	3,467	3,100	2,858	2,685	2,553	2,450	2,366	2,296	2,237	2,186	2,142	2,103	2,069	2,038	2,010
90	5,610	4,087	3,445	3,079	2,837	2,664	2,533	2,429	2,345	2,275	2,216	2,165	2,121	2,082	2,048	2,017	1,989
00	5,590	4,069	3,428	3,062	2,821	2,648	2,517	2,413	2,329	2,259	2,200	2,149	2,105	2,066	2,031	2,000	1,972
∞	5,412	3,912	3,279	2,917	2,678	2,506	2,375	2,271	2,187	2,116	2,056	2,005	1,959	1,919	1,884	1,852	1,823

Table de F (seuil 2 %) [suite]

ddl(B) \ ddl(A)	18	19	20	25	30	35	40	45	50	60	70	80	90	100	∞
1	1547.6	1549.8	1551.9	1559.7	1564.9	1568.6	1571.4	1573.6	1575.3	1578.0	1579.9	1581.3	1582.4	1583.2	1591.2
2	49.443	49.446	49.448	49.458	49.466	49.469	49.473	49.477	49.478	49.482	49.484	49.486	49.487	49.488	49.498
3	16.588	16.570	16,553	16.490	16.448	16.417	16.394	16,376	16,362	16.340	16.325	16.313	16.304	16.297	16.231
4	9.708	9.688	9,670	9,601	9,554	9.521	9.495	9,476	9.460	9,436	9.419	9,406	9,396	9,388	9.315
5	7.048	7.028	7.009	6.940	6.893	6.859	6.833	6.813	6.797	6.773	6.755	6.742	6.732	6.724	6.650
6	5,689	5,669	5,651	5,581	5,534	5,500	5,474	5,454	5,438	5,413	5,396	5,383	5,372	5,364	5.289
7	4.878	4.857	4,839	4,769	4.722	4.688	4,662	4.641	4.625	4.601	4.583	4.570	4.559	4.551	4.475
8	4.342	4.322	4,304	4,233	4,186	4,151	4.125	4,105	4,088	4,063	4.046	4.032	4,022	4.013	3.936
9	3.963	3.943	3.925	3.854	3.806	3.772	3,745	3.725	3,708	3.683	3.665	3.651	3.641	3.632	3.554
10	3.682	3.662	3.644	3.573	3.525	3.490	3.463	3.442	3.425	3.400	3.382	3.368	3.357	3.348	3.269
11	3.466	3.445	3.427	3.355	3,307	3,272	3.245	3,224	3,207	3,181	3.163	3.149	3.138	3.129	3.048
12	3.293	3.273	3.254	3,183	3,134	3.098	3.071	3.050	3.033	3.007	2.988	2.974	2.963	2.954	2.872
13	3.153	3.133	3,114	3.042	2.993	2.957	2.930	2.908	2.891	2.865	2.846	2.831	2.820	2.811	2.728
14	3.037	3.017	2.998	2.926	2.876	2.840	2.812	2.791	2.773	2.747	2.727	2.713	2.702	2.692	2.608
15	2.940	2.919	2.900	2.827	2.777	2.741	2.713	2.691	2.674	2.647	2.627	2.613	2.601	2.592	2.506
16	2,856	2.836	2.817	2,744	2.693	2.656	2,628	2.606	2,589	2.561	2.542	2,527	2.515	2.506	2,419
17	2.784	2.764	2,745	2.671	2.620	2.583	2.555	2.533	2.515	2.487	2.468	2.453	2.441	2.431	2.343
18	2.722	2.701	2.682	2.608	2.557	2.519	2.491	2.469	2.450	2.423	2.403	2.388	2.376	2.366	2.277
19	2.666	2.645	2,626	2.552	2,501	2,463	2.434	2.412	2.394	2.366	2.345	2.330	2.318	2.308	2.218
20	2,617	2.596	2.577	2.502	2.451	2.413	2.384	2.361	2.343	2.315	2.294	2.279	2.267	2.257	2.165
25	2.437	2.416	2,396	2,320	2,267	2,228	2,199	2.175	2.156	2.127	2.105	2.089	2.076	2.066	1.969
30	2.322	2.300	2.281	2,203	2,149	2.110	2.079	2.055	2.035	2.005	1.983	1.966	1.953	1.942	1.840
35	2.242	2.220	2.200	2,122	2,067	2.027	1.995	1.971	1,950	1,919	1.896	1.879	1.865	1.854	1.748
40	2.183	2.161	2.141	2.062	2.006	1.965	1.933	1.908	1.887	1.856	1.832	1.814	1.800	1.789	1.678
45	2.138	2.116	2.096	2,016	1.960	1.918	1.886	1.860	1.839	1.806	1.783	1.764	1.750	1.738	1.623
50	2.103	2.080	2.060	1.979	1.923	1.881	1.848	1.822	1.800	1.767	1.743	1.724	1.709	1.697	1.579
60	2.050	2.027	2.007	1.925	1.868	1.825	1.791	1.765	1.743	1.709	1.683	1.664	1.649	1.636	1.511
70	2.013	1.990	1.969	1.887	1.829	1.785	1.751	1.724	1.701	1.667	1.641	1.621	1.605	1.592	1.462
80	1.985	1.962	1.941	1.858	1.800	1.755	1.721	1.693	1.671	1.635	1.609	1.588	1.572	1.559	1.423
90	1.964	1.941	1.920	1.836	1.777	1.733	1.698	1.670	1.647	1.611	1.584	1.563	1.546	1.533	1.392
100	1.947	1.924	1.902	1.819	1.759	1.714	1.679	1.651	1.627	1.591	1.564	1.543	1.526	1.512	1.367
∞	1.797	1.773	1.751	1.663	1.599	1.550	1.511	1.479	1.452	1.410	1.377	1.351	1.329	1.311	1.000

Table de F (seuil 1 %)

ddl(B) \ ddl(A)	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
1	4052,2	4999,3	5403,5	5624,3	5764,0	5859,0	5928,3	5981,0	6022,4	6055,9	6083,4	6106,7	6125,8	6143,0	6157,0	6170,0	6181,2
2	98,50	99,00	99,16	99,25	99,30	99,33	99,36	99,38	99,39	99,40	99,41	99,42	99,42	99,43	99,43	99,44	99,44
3	34,116	30,816	29,457	28,710	28,237	27,911	27,671	27,489	27,345	27,228	27,132	27,052	26,983	26,924	26,872	26,826	26,786
4	21,198	18,000	16,694	15,977	15,522	15,207	14,976	14,799	14,659	14,546	14,452	14,374	14,306	14,249	14,198	14,154	14,114
5	16,258	13,274	12,060	11,392	10,967	10,672	10,456	10,289	10,158	10,051	9,963	9,888	9,825	9,770	9,722	9,680	9,643
6	13,745	10,925	9,780	9,148	8,746	8,466	8,260	8,102	7,976	7,874	7,790	7,718	7,657	7,605	7,559	7,519	7,483
7	12,246	9,547	8,451	7,847	7,460	7,191	6,993	6,840	6,719	6,620	6,538	6,469	6,410	6,359	6,314	6,275	6,240
8	11,259	8,649	7,591	7,006	6,632	6,371	6,178	6,029	5,911	5,814	5,734	5,667	5,609	5,559	5,515	5,477	5,442
9	10,562	8,022	6,992	6,422	6,057	5,802	5,613	5,467	5,351	5,257	5,178	5,111	5,055	5,005	4,962	4,924	4,890
10	10,044	7,559	6,552	5,994	5,636	5,386	5,200	5,057	4,942	4,849	4,772	4,706	4,650	4,601	4,558	4,520	4,487
11	9,646	7,206	6,217	5,668	5,316	5,069	4,886	4,744	4,632	4,539	4,462	4,397	4,342	4,293	4,251	4,213	4,180
12	9,330	6,927	5,953	5,412	5,064	4,821	4,640	4,499	4,388	4,296	4,220	4,155	4,100	4,052	4,010	3,972	3,939
13	9,074	6,701	5,739	5,205	4,862	4,620	4,441	4,302	4,191	4,100	4,025	3,960	3,905	3,857	3,815	3,778	3,745
14	8,862	6,515	5,564	5,035	4,695	4,456	4,278	4,140	4,030	3,939	3,864	3,800	3,745	3,698	3,656	3,619	3,586
15	8,683	6,359	5,417	4,893	4,556	4,318	4,142	4,004	3,895	3,805	3,730	3,666	3,612	3,564	3,522	3,485	3,452
16	8,531	6,226	5,292	4,773	4,437	4,202	4,026	3,890	3,780	3,691	3,616	3,553	3,498	3,451	3,409	3,372	3,339
17	8,400	6,112	5,185	4,669	4,336	4,101	3,927	3,791	3,682	3,593	3,518	3,455	3,401	3,353	3,312	3,275	3,242
18	8,285	6,013	5,092	4,579	4,248	4,015	3,841	3,705	3,597	3,508	3,434	3,371	3,316	3,269	3,227	3,190	3,158
19	8,185	5,926	5,010	4,500	4,171	3,939	3,765	3,631	3,523	3,434	3,360	3,297	3,242	3,195	3,153	3,116	3,084
20	8,096	5,849	4,938	4,431	4,103	3,871	3,699	3,564	3,457	3,368	3,294	3,231	3,177	3,130	3,088	3,051	3,018
25	7,770	5,568	4,675	4,177	3,855	3,627	3,457	3,324	3,217	3,129	3,056	2,993	2,939	2,892	2,850	2,813	2,780
30	7,562	5,390	4,510	4,018	3,699	3,473	3,305	3,173	3,067	2,979	2,906	2,843	2,789	2,742	2,700	2,663	2,630
35	7,419	5,268	4,396	3,908	3,592	3,368	3,200	3,069	2,963	2,876	2,803	2,740	2,686	2,639	2,597	2,560	2,527
40	7,314	5,178	4,313	3,828	3,514	3,291	3,124	2,993	2,888	2,801	2,727	2,665	2,611	2,563	2,522	2,484	2,451
45	7,234	5,110	4,249	3,767	3,454	3,232	3,066	2,935	2,830	2,743	2,670	2,608	2,553	2,506	2,464	2,427	2,393
50	7,171	5,057	4,199	3,720	3,408	3,186	3,020	2,890	2,785	2,698	2,625	2,563	2,508	2,461	2,419	2,382	2,348
60	7,077	4,977	4,126	3,649	3,339	3,119	2,953	2,823	2,718	2,632	2,559	2,496	2,442	2,394	2,352	2,315	2,281
70	7,011	4,922	4,074	3,600	3,291	3,071	2,906	2,777	2,672	2,585	2,512	2,450	2,395	2,348	2,306	2,268	2,234
80	6,963	4,881	4,036	3,563	3,255	3,036	2,871	2,742	2,637	2,551	2,478	2,415	2,361	2,313	2,271	2,233	2,199
90	6,925	4,849	4,007	3,535	3,228	3,009	2,845	2,715	2,611	2,524	2,451	2,389	2,334	2,286	2,244	2,206	2,172
100	6,895	4,824	3,984	3,513	3,206	2,988	2,823	2,694	2,590	2,503	2,430	2,368	2,313	2,265	2,223	2,185	2,151
∞	6,635	4,605	3,782	3,319	3,017	2,802	2,639	2,511	2,407	2,321	2,248	2,185	2,130	2,082	2,039	2,000	1,965

Table de F (seuil 1 %) [suite]

ddl(B) \ ddl(A)	18	19	20	25	30	35	40	45	50	60	70	80	90	100	∞
1	6191,4	6200,7	6208,7	6239,9	6260,4	6275,3	6286,4	6295,7	6302,3	6313,0	6320,9	6326,5	6330,7	6333,9	6365,6
2	99,44	99,45	99,45	99,46	99,47	99,47	99,48	99,48	99,48	99,48	99,48	99,48	99,49	99,49	99,50
3	26,751	26,719	26,690	26,579	26,504	26,451	26,411	26,379	26,354	26,316	26,289	26,269	26,253	26,241	26,125
4	14,079	14,048	14,019	13,911	13,838	13,785	13,745	13,714	13,690	13,652	13,626	13,605	13,590	13,577	13,463
5	9,609	9,580	9,553	9,449	9,379	9,329	9,291	9,262	9,238	9,202	9,176	9,157	9,142	9,130	9,020
6	7,451	7,422	7,396	7,296	7,229	7,180	7,143	7,115	7,091	7,057	7,032	7,013	6,998	6,987	6,880
7	6,209	6,181	6,155	6,058	5,992	5,944	5,908	5,880	5,858	5,824	5,799	5,781	5,766	5,755	5,650
8	5,412	5,384	5,359	5,263	5,198	5,151	5,116	5,088	5,065	5,032	5,007	4,989	4,975	4,963	4,859
9	4,860	4,833	4,808	4,713	4,649	4,602	4,567	4,539	4,517	4,483	4,459	4,441	4,426	4,415	4,311
10	4,457	4,430	4,405	4,311	4,247	4,201	4,165	4,138	4,115	4,082	4,058	4,039	4,025	4,014	3,909
11	4,150	4,123	4,099	4,005	3,941	3,895	3,860	3,832	3,810	3,776	3,752	3,734	3,719	3,708	3,602
12	3,910	3,883	3,858	3,765	3,701	3,654	3,619	3,592	3,569	3,535	3,511	3,493	3,478	3,467	3,361
13	3,716	3,689	3,665	3,571	3,507	3,461	3,425	3,398	3,375	3,341	3,317	3,298	3,284	3,272	3,165
14	3,556	3,529	3,505	3,412	3,348	3,301	3,266	3,238	3,215	3,181	3,157	3,138	3,124	3,112	3,004
15	3,423	3,396	3,372	3,278	3,214	3,167	3,132	3,104	3,081	3,047	3,022	3,004	2,989	2,977	2,868
16	3,310	3,283	3,259	3,165	3,101	3,054	3,018	2,990	2,967	2,933	2,908	2,889	2,875	2,863	2,753
17	3,212	3,186	3,162	3,068	3,003	2,956	2,920	2,892	2,869	2,835	2,810	2,791	2,776	2,764	2,653
18	3,128	3,101	3,077	2,983	2,919	2,871	2,835	2,807	2,784	2,749	2,724	2,705	2,690	2,678	2,566
19	3,054	3,027	3,003	2,909	2,844	2,797	2,761	2,732	2,709	2,674	2,649	2,630	2,614	2,602	2,489
20	2,989	2,962	2,938	2,843	2,778	2,731	2,695	2,666	2,643	2,608	2,582	2,563	2,548	2,535	2,421
25	2,751	2,724	2,699	2,604	2,538	2,490	2,453	2,424	2,400	2,364	2,337	2,317	2,302	2,289	2,169
30	2,600	2,573	2,549	2,453	2,386	2,337	2,299	2,269	2,245	2,208	2,181	2,160	2,144	2,131	2,006
35	2,497	2,470	2,445	2,348	2,281	2,231	2,193	2,162	2,137	2,099	2,072	2,050	2,034	2,020	1,891
40	2,421	2,394	2,369	2,271	2,203	2,153	2,114	2,083	2,058	2,019	1,991	1,969	1,952	1,938	1,805
45	2,363	2,336	2,311	2,213	2,144	2,093	2,054	2,023	1,997	1,958	1,929	1,907	1,889	1,875	1,737
50	2,318	2,290	2,265	2,167	2,098	2,046	2,007	1,975	1,949	1,909	1,880	1,857	1,839	1,825	1,683
60	2,251	2,223	2,198	2,098	2,028	1,976	1,936	1,904	1,877	1,836	1,806	1,783	1,764	1,749	1,601
70	2,204	2,176	2,150	2,050	1,980	1,927	1,886	1,853	1,826	1,785	1,754	1,730	1,711	1,695	1,540
80	2,169	2,141	2,115	2,015	1,944	1,890	1,849	1,816	1,788	1,746	1,714	1,690	1,671	1,655	1,494
90	2,142	2,114	2,088	1,987	1,916	1,862	1,820	1,787	1,759	1,716	1,684	1,659	1,639	1,623	1,457
100	2,120	2,092	2,067	1,965	1,893	1,839	1,797	1,763	1,735	1,692	1,659	1,634	1,614	1,598	1,427
∞	1,934	1,905	1,878	1,773	1,696	1,638	1,592	1,555	1,523	1,473	1,435	1,404	1,379	1,358	1,000

Table de F (seuil 0,1 %)

ddl(B) \ ddl(A)	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
1	405312	499725	540257	562668	576496	586033	593185	597954	602245	605583	608444	610352	612259	614166	616074	617027	617981
2	998,38	998,84	999,31	999,31	999,31	999,31	999,31	999,31	999,31	999,31	999,31	999,31	999,31	999,31	999,31	999.31	999,31
3	167,06	148,49	141.10	137,08	134,58	132.83	131.61	130,62	129,86	129,22	128,76	128,32	127.94	127,65	127,36	127.13	126,95
4	74,127	61,249	56,170	53,435	51,718	50.524	49,651	48.996	48,472	48.050	47,701	47,410	47,163	46,944	46.755	46.595	46.450
5	47,177	37,122	33,200	31,083	29,751	28.835	28.165	27,649	27,241	26,914	26,645	26,419	26,223	26,055	25.910	25.782	25.670
6	35.507	27.001	23.705	21.922	20,802	20.031	19.463	19.030	18.688	18.412	18,183	17.990	17.826	17.684	17.557	17.451	17.353
7	29,246	21,690	18.772	17.197	16.207	15,520	15,018	14,634	14,330	14,083	13,879	13.708	13.561	13.435	13.324	13.228	13.140
8	25,415	18.494	15.829	14.392	13,484	12.858	12.398	12.045	11.767	11.540	11.352	11.194	11.059	10.943	10.841	10.752	10.672
9	22.857	16,387	13.901	12.560	11.714	11.129	10,697	10.368	10,106	9.894	9.719	9,570	9.443	9.333	9.239	9.153	9.079
10	21.038	14,905	12.553	11.283	10.481	9.926	9.517	9.204	8.956	8.754	8.587	8.446	8.325	8.220	8.129	8.048	7.977
11	19,687	13,812	11,561	10,346	9,579	9,047	8,655	8.355	8,116	7,923	7,762	7,625	7,510	7,409	7,321	7,243	7.175
12	18.645	12,973	10,805	9,633	8,892	8,378	8.001	7,711	7,480	7,292	7,136	7,005	6,892	6.795	6.709	6.634	6,567
13	17.815	12,313	10,209	9,073	8,355	7,856	7,489	7,206	6,982	6,799	6,647	6.519	6,409	6,315	6.231	6,158	6.093
14	17.142	11,779	9,730	8.622	7,922	7,436	7,078	6,802	6,583	6,404	6,256	6,130	6,023	5,930	5.848	5.776	5.713
15	16.587	11.340	9.335	8,253	7,567	7,091	6,741	6.471	6,256	6.081	5,935	5.812	5,707	5.615	5.535	5.464	5.401
16	16.120	10.970	9.006	7.944	7,272	6,805	6,460	6.195	5,984	5.812	5,668	5.547	5.443	5,353	5,275	5.205	5.143
17	15,722	10.658	8.727	7,683	7.022	6,562	6.224	5,962	5.754	5.584	5.443	5.324	5,221	5.132	5.055	4.986	4.924
18	15,380	10.390	8.487	7,460	6,808	6.355	6,021	5.763	5.557	5,390	5,251	5.132	5,031	4.943	4.866	4,798	4.738
19	15.081	10,157	8.280	7,265	6,622	6,175	5.845	5.591	5.387	5.222	5.084	4.967	4.867	4.780	4.703	4.636	4.576
20	14.819	9,953	8.098	7.096	6.461	6.019	5.692	5.440	5.239	5.075	4.939	4.823	4.723	4.638	4.562	4.495	4.435
25	13.877	9.222	7.451	6,493	5.885	5.462	5.148	4.906	4.713	4.555	4,423	4.311	4,216	4,132	4.059	3.994	3.936
30	13,293	8,773	7,054	6.125	5.534	5.122	4.817	4.582	4.393	4.239	4.110	4.001	3.907	3.825	3.753	3.689	3.632
35	12.897	8,470	6,787	5,876	5,298	4,894	4,595	4.363	4,178	4,027	3,900	3.792	3,699	3.618	3.547	3,485	3.428
40	12.609	8.251	6.595	5.698	5,128	4.731	4.436	4.207	4.024	3.874	3.749	3.643	3.551	3.471	3.400	3.338	3.282
45	12,393	8,085	6,450	5,564	5,001	4,608	4,316	4,090	3,909	3,760	3,636	3,530	3,439	3,360	3,290	3.228	3,172
50	12.222	7,956	6.336	5.459	4.901	4,512	4.222	3.998	3.819	3.671	3.548	3,443	3.352	3.273	3,203	3.142	3.086
60	11,973	7,768	6,171	5,307	4,757	4,372	4,086	3,865	3,687	3,542	3,419	3,315	3,226	3,147	3,078	3.017	2.962
70	11.800	7,637	6.056	5,201	4,656	4,275	3.992	3.773	3.596	3.452	3.330	3.227	3,138	3.060	2.991	2.930	2.875
80	11,672	7,540	5,972	5,123	4,582	4,204	3.923	3,705	3,530	3.386	3,265	3,162	3,074	2.996	2.927	2.866	2.812
90	11.573	7.466	5,908	5.064	4.526	4,150	3.870	3.653	3,479	3.336	3.215	3.113	3.024	2.947	2.879	2.818	2.763
100	11.496	7.408	5.857	5.017	4.482	4.107	3.829	3.612	3.439	3.296	3.176	3.074	2.986	2.908	2.840	2.780	2.725
∞	10.828	6.908	5.422	4.617	4.103	3.743	3.474	3.266	3.098	2.959	2.842	2.742	2.656	2.580	2.513	2.453	2.399

Table de F (seuil 0,1 %) [suite]

ddl(B) \ ddl(A)	18	19	20	25	30	35	40	45	50	60	70	80	90	100	∞
1	618935	619888	620842	623703	626087	627518	628471	629425	630379	631332	632286	632286	633240	633240	636578
2	999,31	999,31	999,31	999,31	999,31	999,31	999,31	999,31	999,31	999,31	999,31	999,31	999,31	999,31	999,31
3	126,72	126,57	126,43	125,85	125,44	125,15	124,97	124,80	124,68	124,45	124,33	124,22	124,16	124,07	123,46
4	46,319	46,202	46,100	45,693	45,431	45,235	45,082	44,980	44,878	44,747	44,645	44,573	44,514	44,471	44,049
5	25,568	25,477	25,393	25,080	24,869	24,716	24,600	24,513	24,440	24,331	24,258	24,200	24,149	24,113	23,785
6	17,266	17,189	17,120	16,851	16,673	16,542	16,444	16,367	16,305	16,214	16,149	16,098	16,058	16,029	15,745
7	13,064	12,995	12,931	12,693	12,529	12,413	12,325	12,256	12,202	12,118	12,060	12,014	11,980	11,951	11,696
8	10,601	10,537	10,479	10,259	10,108	10,001	9,919	9,855	9,804	9,728	9,672	9,630	9,597	9,572	9,333
9	9,012	8,952	8,898	8,689	8,547	8,446	8,368	8,308	8,260	8,186	8,135	8,095	8,064	8,038	7,813
10	7,913	7,856	7,803	7,604	7,469	7,371	7,297	7,240	7,192	7,122	7,072	7,034	7,004	6,980	6,762
11	7,113	7,058	7,008	6,815	6,684	6,589	6,517	6,462	6,416	6,348	6,299	6,262	6,234	6,210	5,998
12	6,507	6,454	6,405	6,217	6,090	5,998	5,928	5,873	5,829	5,763	5,714	5,678	5,650	5,627	5,420
13	6,034	5,982	5,934	5,751	5,626	5,535	5,467	5,413	5,370	5,305	5,257	5,222	5,194	5,172	4,967
14	5,655	5,603	5,557	5,377	5,254	5,165	5,098	5,045	5,002	4,938	4,891	4,856	4,829	4,807	4,604
15	5,345	5,294	5,249	5,071	4,950	4,863	4,796	4,743	4,702	4,638	4,592	4,557	4,530	4,508	4,307
16	5,087	5,037	4,992	4,817	4,697	4,611	4,545	4,493	4,451	4,388	4,342	4,308	4,281	4,259	4,059
17	4,869	4,820	4,775	4,602	4,484	4,398	4,332	4,281	4,240	4,177	4,131	4,097	4,070	4,049	3,849
18	4,683	4,634	4,590	4,418	4,301	4,216	4,151	4,100	4,059	3,996	3,951	3,917	3,890	3,869	3,670
19	4,522	4,474	4,430	4,259	4,143	4,058	3,994	3,943	3,902	3,840	3,795	3,761	3,734	3,713	3,514
20	4,382	4,334	4,290	4,121	4,005	3,921	3,856	3,806	3,765	3,703	3,658	3,624	3,598	3,576	3,378
25	3,884	3,837	3,794	3,629	3,515	3,432	3,369	3,319	3,279	3,217	3,173	3,139	3,112	3,090	2,890
30	3,581	3,535	3,493	3,330	3,217	3,135	3,072	3,022	2,981	2,920	2,875	2,841	2,814	2,792	2,589
35	3,378	3,332	3,290	3,128	3,016	2,934	2,871	2,821	2,781	2,719	2,673	2,639	2,612	2,590	2,383
40	3,232	3,186	3,145	2,984	2,872	2,790	2,727	2,677	2,636	2,574	2,528	2,493	2,466	2,444	2,233
45	3,122	3,077	3,036	2,875	2,763	2,681	2,618	2,568	2,527	2,464	2,418	2,383	2,355	2,333	2,118
50	3,037	2,992	2,951	2,790	2,679	2,596	2,533	2,482	2,441	2,378	2,332	2,296	2,269	2,246	2,026
60	2,912	2,867	2,826	2,667	2,555	2,472	2,409	2,358	2,316	2,252	2,205	2,169	2,141	2,118	1,890
70	2,826	2,781	2,741	2,581	2,469	2,386	2,322	2,271	2,229	2,164	2,117	2,080	2,051	2,027	1,793
80	2,763	2,718	2,677	2,518	2,406	2,323	2,258	2,207	2,164	2,099	2,051	2,014	1,984	1,960	1,720
90	2,714	2,670	2,629	2,469	2,357	2,274	2,209	2,157	2,115	2,049	2,000	1,963	1,933	1,909	1,662
100	2,676	2,632	2,591	2,431	2,319	2,235	2,170	2,118	2,076	2,009	1,960	1,922	1,892	1,867	1,615
∞	2,351	2,306	2,266	2,105	1,990	1,903	1,835	1,780	1,733	1,660	1,605	1,561	1,525	1,495	1,000

Table de Kolmogorov-Smirnov dans le cas de deux échantillons de petite taille

N	Cas unilatéral		Cas bilatéral	
	$\alpha = .05$	$\alpha = .01$	$\alpha = .05$	$\alpha = .01$
3	3	—	—	—
4	4	—	4	—
5	4	5	5	5
6	5	6	5	6
7	5	6	6	6
8	5	6	6	7
9	6	7	6	7
10	6-	7	7	8
11	6	8	7	8
12	6	8	7	8
13	7	8	7	9
14	7	8	8	9
15	7	9	8	9
16	7	9	8	10
17	8	9	8	10
18	8	10	9	10
19	8	10	9	10
20	8	10	9	11
21	8	10	9	11
22	9	11	9	11
23	9	11	10	11
24	9	11	10	12
25	9	11	10	12
26	9	11	10	12
27	9	12	10	12
28	10	12	11	13
29	10	12	11	13
30	10	12	11	13
35	11	13	12	
40	11	14	13	

A partir de Goodman, L.A. 1954. Kolmogorov-Smirnov tests for psychological research. Psychol. Bull; 51, 167 et de Massey, F.J. 1951. The distribution of the maximum deviation between two sample cumulative strep functions. Ann Math Statist., *22, 126-127.*

Table de Kolmogorov-Smirnov dans le cas de deux échantillons de grande taille

Si $|C_C| > C_T$, la différence, ou l'inégalité, est significative entre les deux lois.

C_T	α (test unilatéral)	α (test bilatéral)
1,22	0,05	0,10
1,36	0,025	0,05
1,48	0.0125	0,025
1,63	0,005	0,01
1,73	0,0025	0,005
1,95	0.0005	0,001

D'après Beyer WH, in Handbook of tables for probability and statistics. CRC Press ; 1985.

Table du test de Mann et Whitney, pour $\alpha \leq 5$ %

La table donne la limite inférieure de U au seuil 5 %. U est la plus petite des deux valeurs Uxy ou Uyx avec n_1 et n_2 étant les effectifs des deux séries (n_1 étant le plus petit). Les cases vides (-) sont celles pour lesquelles la différence n'est jamais significative au seuil $\alpha \leq 5$ %.

n_2-n_1 \ n_1	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
0	-	-	-	0	2	5	8	13	17	23	30	37	45	55	64	75	87	99	113	127
1	-	-	-	1	3	6	10	15	20	26	33	41	50	59	70	81	93	106	119	
2	-	-	0	2	5	8	12	17	23	29	37	45	54	64	75	86	99	112		
3	-	-	1	3	6	10	14	19	26	33	40	49	59	69	80	92	105			
4	-	-	1	4	7	11	16	22	28	36	44	53	63	74	85	98				
5	-	2	4	8	13	18	24	31	39	47	57	67	78	90						
6	-	0	2	5	9	14	20	26	34	42	51	61	72	83						
7	-	0	3	6	11	16	22	29	37	45	55	65	76							
8	-	0	3	7	12	17	24	31	39	48	58	69								
9	-	0	4	8	13	19	26	34	42	52	62									
10	-	1	4	9	14	21	28	36	45	55										
11	-	1	5	10	15	22	30	38	48											
12	-	1	5	11	17	24	32	41	50											
13	-	1	6	11	18	25	34	43												
14	-	1	6	12	19	27	36	45												
15	-	2	7	13	20	29	38													
16	-	2	7	14	22	30	40													
17	-	2	8	15	23	32														
18	-	2	8	16	24	33														
19	-	3	9	17	25															
20	-	3	9	17																
21	-	3	10	18																
22	-	3	10	19																
23	-	3	11																	
24	-	4	11																	
25	-	4																		
26	-	4																		

Exemple : Pour $n_1 = 8$, $n_2 = 15$, soit $n_2-n_1 = 7$, la différence est significative au seuil $\alpha \leq 5$ % dès que la statistique U calculée est inférieure à U de la table, soit $U_{réf} = 29$.

D'après JE Jacobson, J. Amer. Statis. Assoc., 1963, 1086.

Table du test de Mann et Whitney, pour $\alpha \leq 1\%$

La table donne la limite inférieure de U au seuil 1 %.
U est la plus petite des deux valeurs Uxy ou Uyx avec n_1 et n_2 étant les effectifs des deux séries (n_1 étant le plus petit). Les cases vides (-) sont celles pour lesquelles la différence n'est jamais significative au seuil $\alpha \leq 1$ %.

n_2-n_1 \ n_1	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
0	-	-	-	-	0	2	4	7	11	16	21	27	34	42	51	60	70	81	93	105
1	-	-	-	-	1	3	6	9	13	18	24	31	38	46	55	65	75	87	99	
2	-	-	-	0	1	4	7	11	16	21	27	34	42	50	60	70	81	92		
3	-	-	-	0	2	5	9	13	18	24	30	37	45	54	64	74	86			
4	-	-	-	1	3	6	10	15	20	26	33	41	49	58	69	79				
5	-	-	-	1	4	7	12	17	22	29	36	44	53	63	73					
6	-	-	0	2	5	9	13	18	24	31	39	47	57	67						
7	-	-	0	2	6	10	15	20	27	34	42	51	60							
8	-	-	0	3	7	11	16	22	29	37	45	54								
9	-	-	1	3	7	12	18	24	31	39	48									
10	-	-	1	4	8	13	19	26	33	42										
11	-	-	1	5	9	15	21	28	36											
12	-	-	2	5	10	16	22	30	38											
13	-	-	2	6	11	17	24	32												
14	-	-	2	6	12	18	25	34												
15	-	-	2	7	13	19	27													
16	-	-	3	8	14	21	29													
17	-	0	3	8	14	22														
18	-	0	3	9	15	23														
19	-	0	4	9	16															
20	-	0	4	10	17															
21	-	0	4	10																
22	-	0	5	11																
23	-	0	5																	
24	-	0	5																	
25	-	1																		
26	-	1																		

Exemple : Pour $n_1 = 8$, $n_2 = 15$, soit $n_2 - n_1 = 7$, la différence est significative au seuil $\alpha \leq 1$ % dès que la statistique U calculée est inférieure à U de la table, soit $U_{réf} = 20$.

D'après JE Jacobson, J. Amer. Statis. Assoc., 1963, 1086.

Table de Wilcoxon pour les séries appariées

La table donne la limite inférieure de la statistique T de Wilcoxon pour les risques de première espèce α inférieurs à 1 % et 5 %. T est le plus petit des deux totaux M (somme des rangs des différences de signe négatif) ou P (somme des rangs des différences de signe positif) en fonction du nombre n de paires non nulles. La différence n'est jamais significative pour $n < 6$ quand $\alpha \leq 5$ %, elle ne l'est jamais pour $n < 8$ quand $\alpha \leq 1$ %.

n \ α	0,025 0,05	0,005 (Test unilatéral) 0,01 (Test bilatéral)
6	0	-
7	2	-
8	4	0
9	6	2
10	8	3
11	11	5
12	14	7
13	17	10
14	21	13
15	25	16
16	30	20
17	35	23
18	40	28
19	46	32
20	52	38
21	59	43
22	66	49
23	73	55
24	81	61
25	89	68

A partir de la table I de Wilcoxon, F. 1949. Some rapid approximate Statistical procedures. New York. American Cyanamid Company, p. 13

Exemple : Dans le cas bilatéral, pour $n_1 = n_2 = 20$, au seuil $\alpha \leq 5$ %, si $W_C > W_T$, on peut conclure à une différence significative (avec W_C correspondant au W calculé et W_T au W lu dans la table). C'est-à-dire que :
- si $W_C > 52$, on ne rejette pas H_0 ;
- au seuil $\alpha \leq 5$ %, si $W_C \leq 52$, on conclut à une différence significative à 5 % ;
- au seuil $\alpha \leq 1$ %, si $W_C \leq 38$, on conclut à une différence significative à 1 % ($p < 0{,}01$).

Attention, le risque de première espèce α doit être fixé *a priori*. Si $W_C \leq 38$ alors que l'on a fixé comme risque de première espèce $\alpha = 5$ %, alors la différence est significative à 5 % et le degré de signification est de 1 % (dans le cas bilatéral).

Table de Kruskal-Wallis
Analyse de variance à un facteur
Valeur de H dans le cas de 3 échantillons de petite taille

Taille des échantillons			H	α	Taille des échantillons			H	α
n_1	n_2	n_3			n_1	n_2	n_3		
2	1	1	2.7000	.500	4	3	2	6.4444	.008
								6.3000	.011
2	2	1	3.6000	.200				5.4444	.046
								5.4000	.051
2	2	2	4.5714	.067				4.5111	.098
			3.7143	.200				4.4444	.102
3	1	1	3.2000	.300					
					4	3	3	6.7455	.010
3	2	1	4.2857	.100				6.7091	.013
			3.8571	.133				6.7909	.046
								5.7273	.050
3	2	2	5.3672	.029				4.7091	.092
			4.7143	.048				4.7000	.101
			4.5000	.067					
			4.4643	.105	4	4	1	6.6667	.010
								6.1667	.022
3	3	1	5.1429	.043				4.9667	.048
			4.5714	.100				4.8667	.054
			4.0000	.129				4.1667	.082
								4.0667	.102
3	3	2	6.2500	.011					
			5.3611	.032	4	4	2	7.0364	.006
			5.1389	.061				6.8727	.011
			4.5556	.100				5.4545	.046
			4.2500	.121				5.2364	.052
3	3	3	7.2000	.004				4.5545	.098
			6.4889	.011				4.4455	.103
			5.6889	.029					
			5.6000	.050	4	4	3	7.1439	.010
			5.0667	.086				7.1364	.011
			4.6222	.100				5.5958	.049
								5.5758	.051
4	1	1	3.5714	.200				4.5455	.099
								4.4773	.102
4	2	1	4.8214	.057					
			4.5000	.076	4	4	4	7.6538	.008
			4.0179	.114				7.5385	.011
								5.6923	.049
4	2	2	6.0000	.014				5.6538	.054
			5.3333	.033				4.6539	.097
			5.1250	.052				4.5001	.104
			4.4583	.100					
			4.1667	.105	5	1	1	3.8571	.143
4	3	1	5.8333	.021	5	2	1	5.2500	.036
			5.2083	.050				5.0000	.048
			5.0000	.057				4.4500	.071
			4.0556	.093				4.2000	.095
			3.8889	.129				4.0500	.119

Table de Kruskal-Wallis (suite)
Analyse de variance à un facteur
Valeur de H dans le cas de 3 échantillons de petite taille

Taille des échantillons			H	α	Taille des échantillons			H	α
n_1	n_2	n_3			n_1	n_2	n_3		
5	2	2	6.5333	.008				5.6308	.060
			6.1333	.013				4.5487	.099
			5.1600	.034				4.5231	.103
			5.0400	.056					
			4.3733	.090	5	4	4	7.7604	.009
			4.2933	.122				7.7440	.011
								5.6571	.049
5	3	1	6.4000	.012				5.6176	.050
			4.9600	.048				4.6187	.100
			4.8711	.062				4.5527	.102
			4.0178	.095					
			3.8400	.123	5	5	1	7.3091	.009
								6.8364	.011
								5.1273	.046
5	3	2	6.9091	.009				4.9091	.053
			6.8218	.010				4.1091	.086
			5.2509	.049				4.0364	.105
			5.1055	.052					
			4.6509	.091	5	5	2	7.3385	.010
			4.4945	.101				7.2692	.010
								5.3385	.047
5	3	3	7.0788	.009				5.2462	.051
			6.9818	.011				4.6231	.097
			5.6485	.049				4.5077	.100
			5.5152	.051					
			4.5333	.097	5	5	3	7.5780	.010
			4.4121	.109				7.5429	.010
								5.7055	.046
5	4	1	6.9545	.008				5.6264	.051
			6.8400	.011				4.5451	.100
			4.9855	.044				4.5363	.102
			4.8600	.056					
			3.9873	.098	5	5	4	7.8229	.010
			3.9600	.102				7.7914	.010
								5.6657	.049
5	4	2	7.2045	.009				5.6429	.050
			7.1182	.010				4.5229	.099
			5.2727	.049				4.5200	.101
			5.2682	.050					
			4.5409	.098	5	5	5	8.0000	.009
			4.5182	.101				7.9800	.010
								5.7800	.049
5	4	3	7.4449	.010				5.6600	.051
			7.3949	.011				4.5600	.100
			5.6564	.049				4.5000	.102

A partir de Kruskal, W.H., et Wallis, W.A., 1952, use of ranks in one - criterion variance analysis. J. Amer Statist. Ass., *47, 614-617*

Table du coefficient de corrélation ρ de Bravais-Pearson

La table donne la limite supérieure de la valeur absolue de ρ pour les risques α inférieurs à 1 % et 5 % en fonction du nombre n de couples de valeurs.

n \ α	0,10	0,05	0,02	0,001
1	0,9877	0,9969	0,9995	0,9999
2	0,9000	0,9500	0,9800	0,9900
3	0,8054	0,8783	0,9343	0,9587
4	0,7293	0,8114	0,8822	0,9172
5	0,6694	0,7545	0,8329	0,8745
6	0,6215	0,7067	0,7887	0,8343
7	0,5822	0,6664	0,7498	0,7977
8	0,5494	0,6319	0,7155	0,7646
9	0,5214	0,6021	0,6851	0,7348
10	0,4973	0,5760	0,6581	0,7079
11	0,4762	0,5529	0,6339	0,6835
12	0,4575	0,5324	0,6120	0,6614
13	0,4409	0,5139	0,5923	0,6411
14	0,4259	0,4973	0,5742	0,6226
15	0,4124	0,4821	0,5577	0,6055
16	0,4000	0,4683	0,5425	0,5897
17	0,3887	0,4555	0,5285	0,5751
18	0,3783	0,4438	0,5155	0,5614
19	0,3687	0,4329	0,5034	0,5487
20	0,3598	0,4227	0,4921	0,5368
25	0,3233	0,3809	0,4451	0,4869
30	0,2960	0,3494	0,4093	0,4487
35	0,2746	0,3246	0,3810	0,4182
40	0,2573	0,3044	0,3578	0,3932
45	0,2428	0,2875	0,3384	0,3721
50	0,2306	0,2732	0,3218	0,3541
60	0,2108	0,2500	0,2948	0,3248
70	0,1954	0,2319	0,2737	0,3017
80	0,1829	0,2172	0,2565	0,2830
90	0,1726	0,2050	0,2422	0,2673
100	0,1638	0,1946	0,2301	0,2540

D'après Fisher et Yates, Statistical tables for biological, agricultural and medical research (Oliver and Boyd. Edinburgh)

Table du coefficient de corrélation des rangs ρ' de Spearman

La table donne la limite supérieure de la valeur absolue de ρ' pour les risques α inférieurs à 1 % et 5 % en fonction du nombre n de couples de valeurs. La différence n'est jamais significative pour $n < 5$ quand $\alpha \leq 5$ %, elle ne l'est jamais pour $n < 6$ quand $\alpha \leq 1$ %.

n \ α	0,05	0,01
5	1,00	-
6	0,89	1.00
7	0,79	0,93
8	0,74	0,88
9	0.68	0,83
10	0,65	0,79

LA COLLECTION "PSYCHO"

Ouvrages déjà parus :

- ***L'entretien en clinique***
  sous la direction de Catherine Cyssau
  (345 pages, 120 F)

- ***La psychologie projective en pratique professionnelle***
  sous la direction de Valérie Boucherat-Hue
  (265 pages, 110 F)

- ***Histoire de la psychologie générale***
  Du behaviorisme au cognitivisme
  Claudette Mariné et Christian Escribe
  (200 pages, 110 F)

- ***Psychologie cognitive***
  Bernard Cadet
  (345 pages, 130 F)

LA COLLECTION *Psycho*

- Une collection pour les étudiants en psychologie, les psychologues, les psychiatres, les professionnels de la santé et de l'éducation.

- Une collection de référence pour les études de psychologie adaptée à tous les cursus quels que soient le cycle ou la Faculté.

- De la psychologie clinique à la psychologie cognitive, des statistiques en psychologie à l'histoire de la psychologie..., toutes les approches qui fondent cette discipline sont traitées – avec clarté et rigueur – par des universitaires reconnus.

Statistique en psychologie

Nathalie Rude et Olivier Retel

Les statistiques sont aujourd'hui un outil indispensable en psychologie : pour vérifier des hypothèses, étudier des liens entre variables, évaluer des conduites aléatoires... quelle que soit la discipline – en psychologie sociale, différentielle, cognitive, en psychologie du travail, de la santé ou du développement. Pourtant, elles sont souvent perçues comme un obstacle, une « barrière de chiffres » difficile à franchir.

Cet ouvrage clarifie les concepts, met en évidence la finalité et l'utilité des statistiques, dissipe les appréhensions. Simple et accessible, il permet de comprendre les principes nécessaires à la réalisation d'analyses statistiques, quelle que soit la formation initiale – scientifique ou littéraire – du lecteur.

Clair, pratique, complet, il couvre l'ensemble du programme des statistiques en psychologie. Les principaux tests y sont présentés. Rôle et principe du test, procédures et conditions d'application sont détaillés. Chacun des tests est illustré par un ou plusieurs exemples qui permettent d'appliquer les principes à des cas concrets issus du domaine de la psychologie. Enfin, dans chaque chapitre, des exercices aident à tester ses connaissances et à se préparer aux examens. Véritable outil d'étude et de travail, cet ouvrage met les méthodes statistiques à la portée de chacun.

ISBN 2-912404-08-8 140 F

9 782912 404084

Conception : Clarisse Enée
Maquette : Marianne Adato

In press

LA COLLECTION *Psycho*

- Une collection pour les étudiants en psychologie, les psychologues, les psychiatres, les professionnels de la santé et de l'éducation.
- Une collection de référence pour les études de psychologie adaptée à tous les cursus quels que soient le cycle ou la Faculté.
- De la psychologie clinique à la psychologie cognitive, des statistiques en psychologie à l'histoire de la psychologie..., toutes les approches qui fondent cette discipline sont traitées – avec clarté et rigueur – par des universitaires reconnus.

Statistique en psychologie

Nathalie Rude et Olivier Retel

Les statistiques sont aujourd'hui un outil indispensable en psychologie : pour vérifier des hypothèses, étudier des liens entre variables, évaluer des conduites aléatoires... quelle que soit la discipline – en psychologie sociale, différentielle, cognitive, en psychologie du travail, de la santé ou du développement. Pourtant, elles sont souvent perçues comme un obstacle, une « barrière de chiffres » difficile à franchir.

Cet ouvrage clarifie les concepts, met en évidence la finalité et l'utilité des statistiques, dissipe les appréhensions. Simple et accessible, il permet de comprendre les principes nécessaires à la réalisation d'analyses statistiques, quelle que soit la formation initiale – scientifique ou littéraire – du lecteur.

Clair, pratique, complet, il couvre l'ensemble du programme des statistiques en psychologie. Les principaux tests y sont présentés. Rôle et principe du test, procédures et conditions d'application sont détaillés. Chacun des tests est illustré par un ou plusieurs exemples qui permettent d'appliquer les principes à des cas concrets issus du domaine de la psychologie. Enfin, dans chaque chapitre, des exercices aident à tester ses connaissances et à se préparer aux examens.

Véritable outil d'étude et de travail, cet ouvrage met les méthodes statistiques à la portée de chacun.

ISBN 2-912404-08-8 140 F

9 782912 404084

Conception : Clarisse Enée
Maquette : Marianne Adato

In press